Advanced Breakdown Modeling for Solid-State Circuit Design

Vladimir Milovanović

Advanced Breakdown Modeling for Solid-State Circuit Design

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen

op woensdag 7 juli 2010 om 10:00 uur

door

Vladimir MILOVANOVIĆ

diplomirani inženjer elektrotehnike
van Univerzitet u Beogradu, Srbija,
geboren te Smederevska Palanka, Srbija

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. L. K. Nanver

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter	Technische Universiteit Delft
Prof. dr. L. K. Nanver	promotor	Technische Universiteit Delft
Dr. ir. R. van der Toorn	copromotor	Technische Universiteit Delft
Prof. dr. ir. J. W. Slotboom		Technische Universiteit Delft
Prof. dr. J. R. Long		Technische Universiteit Delft
Prof. dr. P. Pejović		Univerzitet u Beogradu, Srbija
Dr. D. B. M. Klaassen		NXP Semiconductors, Eindhoven
Dr. J. Victory		Sentinel IC Technologies, California
Prof. dr. E. Charbon	reserve lid	Technische Universiteit Delft

Dr. S. Mijalković heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Vladimir Milovanović,
Advanced Breakdown Modeling for Solid-State Circuit Design,
PhD Thesis, Delft University of Technology.

Keywords: avalanche, breakdown, circuit design, compact model, modeling, p-n junction, parameter extraction, statistical analysis, transistor, tunneling, Zener.

ISBN: 978-90-8570-583-3

Printed by CPI, Wöhrmann Print Service, Zutphen, the Netherlands

Својој породици

Contents

Chapter 1

Introduction

A TENDENCY and will to communicate, together with specific ways of thinking and feeling, are the ones mostly shaping human social nature and human nature in general. Since the evolution of *Homo sapiens* species up to the present day, this wish for mutual exchange of information has not ceased. On the contrary, it only becomes more prominent and pronounced with the time. However, what has changed ever since is the way this communication between humans is carried out.

One of the oldest form of communication in recorded history are certainly smoke signals. In this way, a signaler was able to transmit a message as far as several hundreds of kilometers in just a few hours. Flags and pennants have been used to dispatch messages across shorter distances. Semaphore telegraph was a system of conveying information by means of visual signals, using towers with pivoting shutters. Information is encoded by the position of the mechanical elements. Not only visual communications but the sound ones, drums and horns or loud whistles have also served as an early form of long distance communication. Carrier pigeons were used to carry messages not so long ago. Nonetheless, it is not until the modern age of electricity and electronics that the real telecommunication revolution began.

The first breakthrough into modern electrical telecommunications came with the development of the telegraph. It is followed by the invention of the telephone. The use of these electrical means of communication exploded even into forms for transcontinental communication via cables on the floors of the ocean. The heaviest handicap of these types of communication are the conducting metal wires they require.

Foundation pillars for wireless telecommunications were placed by James Clerk Maxwell and his consistent model of electromagnetism described by the set of four equations. Maxwell demonstrated that electric and magnetic fields travel through space in the form of electromagnetic waves [1] at the speed of light. After the theory was ready, other inventors, including (and led [2] by) Nikola Tesla were there to exploit it, thus assuring another telecommunication revolution, the wireless one.

The existence of theory and experimental setups that work and that may even have commercial applications are still far from everyday consumer products which we all witness at the very moment. A century long period has elapsed from the first working examples to everyday personal use. In this period several epochal technological discoveries had to take place, such as the invention of the transistor, of the integrated circuit, etc. It can be argued that the current success of wireless telecommunication technology with mass-market appeal has been made possible by the cost, size and performance advantages of solid-state semiconductor integration, above all in silicon.

As the trend for higher bit transfer rates and increasing system throughputs continued, theories setting some fundamental limitations on it also emerged. Especially important from the prospect of motivation for this thesis is the Shannon's theorem [3]. It establishes channel capacity C, the theoretical tightest upper bound on the information rate (excluding error correcting codes) of clean (arbitrarily low bit error rate) data/information that can be sent with a given average signal power or power spectral density S through a continuous-time analog communication channel of specific bandwidth B subject to the additive white Gaussian noise of power N, in a form of

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \quad .$$

The theorem yields two options for increasing the channel capacity in the presence of fixed noise. It may be achieved by widening the bandwidth or by increasing the signal to noise ratio (SNR). Boosting transmission power in autonomous mobile systems inevitably leads to tradeoffs with battery life span. Also, the channel capacity linearly increases with bandwidth and logarithmically with SNR, so from that prospective the increase of (passband) bandwidth looks favorable. Logically, (absolutely speaking) more frequency bandwidth is available on higher frequencies. Therefore, the need for higher operating frequencies in order to enlarge and speed-up data rates is justified.

In order to keep the exponential trend of the data rate increase, processes that allow active devices working on ever higher operating frequencies have to be continuously developed. As always, the price is of fundamental importance as for every consumer electronics product, so these solutions have to be as cheap as possible. This draws attention on both process technologists and circuit designers that have to be able to shift paradigms in order to enable further technological progress.

1.1 Monolithic semiconductor solutions

Complementary metal-oxide-semiconductor (CMOS) is already for several decades technology of choice for digital integrated circuit (IC) design, while bipolar junction transistors (BJTs) are traditionally strong when analog circuits are in consideration. Recent acceleration in operating frequency of radio frequency (RF) circuits, which saw a move from centimeter to millimeter wave (mmW) band, is largely enabled by low-cost silicon germanium (carbon) SiGe(:C) heterojunction bipolar transistors (HBTs). Idea to combine these, superb CMOS digital logic performance with bipolar transistors that would be used for analog parts, in order to provide a possibility for designers to exploit the best of the two in mixed-signal integrated circuits, resulted in BiCMOS process technologies. Present day SiGe BiCMOS processes integrate high-performance HBTs with recent CMOS technology. With each generation of the technology, the reduction of feature size leads to higher operating speeds and lower power consumptions of manufactured integrated circuits.

There are several reasons why a SiGe BiCMOS process is advantageous over RF CMOS (that is basically pure CMOS process accommodated to RF needs), for analog circuit designers. As the unity current gain and unity power gain frequencies of both MOSFETs and HBTs can peak in the terahertz [4] range [5] [6, 7, 8] other transistor features and properties emerge as decisive ones for analog RF circuit design. Higher transconductance thanks to intrinsic exponential dependence of the output current on the input voltage of BJTs compared to the square law dependence of FETs [9] and current carrying capability, that is transconductance per bias current, speak in favor of HBTs. For the reason of ever increasing doping concentrations, matching of HBT turn-on voltages is increased, opposite to FETs where ever shrinking lateral dimensions make threshold voltage less stable from generation to generation. Also the optimum source match necessary for proper RF design is more reactive for FETs and more resistive for BJTs, making it much easier to select an optimum matching network for the HBTs and easier to simultaneously optimize both minimum noise figure and associated gain. For comparable cutoff frequency, by cause of higher transconductance, parasitic capacitances in BJTs are considerably higher making the HBT design less affected by layout added parasitics, by that increasing design margin which is good for first pass design and hence faster time to market than RF CMOS. In HBTs, values of $1/f$ noise that are as much as an order of magnitude lower than those in FETs, can be achieved. Finally, larger breakdown voltages and overall superior device-level performance with very few tradeoffs make HBTs of SiGe BiCMOS technology transistors of choice for mixed-signal system-on-a-chip (SoC) solutions.

The key challenges in CMOS analog and RF circuit design is designing high-quality analog circuits with a low transconductance to current ratio [10] and high-performance RF power amplifiers (PAs) with low breakdown voltage. As far as the price is concerned, integrating bipolar transistors within CMOS adds masks and processing steps which necessarily raise the cost of producing the wafer, as a result, for the same feature length, pure and RF CMOS are less costly than the BiCMOS. However, the number of masking steps increases with CMOS scaling and, given the masking cost increase per node, the cost advantage offered by CMOS begins to erode as geometries decrease. Not less important is that although RF CMOS benefits from continued lithography scaling, in terms of performance, it trails SiGe BiCMOS by two generations (technology nodes) [11]. Thereupon, for the same cutoff and maximum oscillation frequency RF CMOS is currently more expensive than the SiGe BiCMOS process.

Still, RF CMOS has impressive RF performance gains with scaling. Cutoff frequency improves with shrinking gate length and thinner gate oxides. A general rule of thumb is that the transition from BiCMOS [12] to RF CMOS comes at a difference of two lithography generations. This means that during the early stages of new application, first generation products will be fabricated in SiGe BiCMOS to leverage both design margin (often critical for time to market) and cost advantages over RF CMOS. As an application matures and becomes more cost sensitive, and CMOS costs come down, following historical trends [13], subsequent generation products will tend to be fabricated with RF CMOS process technology.

A huge advantage of CMOS for digital switching is that the oxide layer between gate and channel prevents DC current from flowing through the gate, thereby reducing power consumption and giving a very large input impedance. This insulating oxide between the gate and channel effectively isolates a metal-oxide-semiconductor field-effect transistor (MOSFET) in one logic stage from earlier and later stages, which allows a single MOSFET output to drive a considerable number of MOSFET inputs [14]. Bipolar transistor-based logic, such as emitter-coupled logic (ECL), does not have such a high fanout capacity but can achieve much higher throughputs [15] essential in optical communication links. Nonetheless, unmatched ability of CMOS technology to reduce average half-pitch size in an exponential rate for the past few decades is certainly one of its incisive and unambiguous benefits.

Silicon versus compound semiconducting materials

The decision to use SiGe BiCMOS is first and foremost an economic one. The ability to combine SiGe HBTs with dense CMOS provides a path to functional integration and gives an obvious advantage over other compound semiconductor solutions such as gallium arsenide (GaAs), gallium nitride (GaN) or indium phosphide (InP). Technical advances in SiGe technology over the past decade [16] have allowed heterojunction bipolar transistor device performance to soar to unexpected heights and openly challenge more expensive III-V compounds. The key advantage of SiGe HBTs is performance close to that of a GaAs HBTs, but implemented in a low-cost silicon process. It has been shown that measured performances of SiGe and GaAs HBT power amplifiers were comparable [17] at cellular bands. SiGe BiCMOS technology has made it possible [18] to address a wide range of applications with silicon-based monolithic solutions. Some of these applications could previously only be addressed with III-V compound semiconductors with low level of integration.

Enabling factors for speed improvement do not fully explain the observed comparable or even superior performance recently achieved by SiGe HBTs compared to that of III-V ones. Benefits of the band gap engineering to be listed below, apply exactly to III-V devices. Furthermore, III-V devices benefit from the well-known material advantages such as higher mobility and more pronounced ballistic carrier transport. In fact, the principal advantage of mainstream Si devices come from their extremely aggressive scaling and extensively optimized structures, that are enabled by state of the art Si technology. Availability of deep submicrometer lithography, precise anisotropic etch, planar structures, high-quality thermal oxide, and silicidation for low contact resistance all contribute to the realization of such aggressively scaled complex device structures. With the operation speed on par, Si-based devices enjoy an additional set of advantages such as CMOS compatibility, large wafer size, and large-scale integration with high yield, all contributing to a strong cost efficiency.

1.2 Breakdown limits of semiconductor devices

The limiting factor of voltage drop over certain section of semiconductor devices is relatively sudden increase of (leakage) current beyond certain point, that can lead to device breakdown if the voltage is further raised. The main physical mechanisms behind the current increase leading to breakdown phenomena are impact ionization that causes avalanche multiplication and valence to conduction band tunneling.

1.2.1 Avalanche breakdown

Scaling has been the key enabler for semiconductor device speed enhancements, but it also brings issues regarding device operation constraints and reliability. Increase of doping concentrations leads to ever higher electric field values within p-n junctions' depletion region which further imply more pronounced impact ionization and avalanche effects, hence reducing device breakdown voltage. Also an increase of doping concentration is often accompanied by an increased operation current density.

Transistor operating speed versus breakdown voltage

The key active device parameters for enhanced circuit performance of noise and linearity for most RF SoC applications are the maximum short-circuit unity current gain frequency f_T and the maximum unity power gain frequency f_max. These two parameters have made astonishing progress in recent years in both HBTs and MOSFETs. High device cutoff frequency, besides being a must for, say, mmW applications, can be traded for other key quantities in the today's volume RF applications that target modest operating frequencies relative to peak f_T. These benefits include [19] reduced power consumption essential for low power applications, higher breakdown voltage that is one of the top priorities in power and voltage amplifiers and reduced noise, of essence in low noise amplifiers (LNAs).

Another absolutely crucial issue is the breakdown voltage, which together with noise considerations influences the dynamic range of operation of most analog (integrated) circuits. The breakdown voltage of a transistor is mostly an issue for the implementation of power amplifiers (PAs) in the (wireless) transmitter section, although other circuit areas can benefit from a high breakdown voltage as well. The breakdown voltage issue is complicated by the physics of the device at high electric fields, the varied physical mechanisms that lead to device failure, and the interaction of the breakdown mechanisms with the external circuit.

Breakdown and degradation in MOS field-effect transistors

Breakdown mechanisms limiting RF MOSFET's performance are complicated by diverse breakdown factors [20], primarily time-dependent dielectric breakdown (TDDB) due to impact ionization in the drain region, gate-oxide rupture, drain avalanche breakdown, parasitic bipolar transistor operation and punchthrough. However, it mainly depends on the gate oxide, which will fail when the applied voltage exceeds the breakdown strength. Also, from a reliability perspective, TDDB presents the most significant limitation on dynamic range in highly scaled MOSFETs. This effect is a result of damage to the silicon oxide interface due to injection of hot carriers (electrons) at the drain [21]. This shifts the threshold voltage of the device over an extended period of time [22]. The recommended voltage limitations are typically based on DC or transient reliability tests, but in many RF applications the instantaneous voltage can significantly exceed the DC voltage, with potentially deleterious consequences. In some cases this phenomenon has been observed to degrade the output power of a CMOS power amplifiers over a matter of days of operation.

Breakdown in (heterojunction) bipolar transistors

A comparison of the HBT breakdown voltages and the recommended operating voltage of MOSFETs as a function of cutoff frequency f_{T} indicates an important conclusion [23]. There seems to be a small but significant advantage in favor of bipolar devices in the high-voltage regime, which is attributed to the fact that there is a cumulative degradation mechanism when the MOSFET is operated in the weak avalanche range of operation (due to long-term shift in the threshold voltage). By comparison, vast majority of bipolar transistors appear to recover without any permanent nor temporary degradation in performance from the weak avalanche breakdown in the base-collector junction. However, in some highly scaled HBTs produced in the latest SiGe processes, the effect of avalanche generated hot carrier oxide damage is visible. It turned out to be more pronounced because emitter and base region thicknesses became fairly low, so that the base-collector metallurgical junction is relatively close to the surface (in vertical bipolar transistors). Notwithstanding, the effect's magnitude is far away from that observed in RF MOSFETs, since the main current flow in bipolar transistors in through the bulk, as opposed to MOS transistors where the current is mainly concentrated in the inversion channel which is close to surface. Effects of hot electrons will have a significant impact on the design of power amplifiers in these technologies, although it should be noted that discrete laterally diffused (LD) MOS devices exhibit excellent performance in high-power base station amplifiers. Nevertheless, they are (still) far behind from use in mmW applications. In case of LDMOS, the devices are engineered to exhibit a very high breakdown voltage as well as acceptable gain at microwave frequencies, which is very different from design considerations that go into typical digital CMOS device scaling. Also it is notable that the collector current density of the fastest Si-based bipolar transistor is still much smaller than the channel current density of typical CMOS devices, implying the margin in the current density increase for Si-based bipolar transistors in terms of bulk Si integrity.

Therefrom, study of pure avalanche multiplication limitation effects seems to be more beneficial for highly attractive BiCMOS HBT devices rather than for pure RF CMOS devices because their performance is limited by other physical mechanisms.

Tradeoffs in bipolar transistors and band gap engineering

As long as the trend in communications progress toward ubiquitous connectivity prolongs, the requirement for higher operational frequencies and increased bandwidth will continue. Transistor current and power gain and thereby bandwidth increase as the output bias voltage is increased (to a certain extent) for the sake of depletion layer width increase and transit time decrease. For bipolar transistors that are used in the critical analog stages of a mixed-signal BiCMOS systems, collector-emitter voltage is pushed to the breakdown limits. Breakdown voltage is important in circuits where either high output power or high output voltage swings are required. A central problem in high speed HBT power stages is the required tradeoff between high output voltage in the weak conducting or off state and low output voltage in the strong conduction state. This means that the designer and technologist have to find an adequate compromise between high breakdown voltage and high current carrying capability of the transistors. The first demand calls for a weakly doped, thick collector region, the second one for a strongly doped, thin region. This problem is most pronounced in high speed silicon technologies due to comparatively low breakdown voltage.

With regard to circuit and device prospective, the tradeoffs in modern HBT device design need to be well understood in order to achieve maximum for a given application. The peak cutoff frequency f_{T}, breakdown voltage (BV) and Early voltage V_{A}, are three parameters that are closely linked in a bipolar transistor. There is a reciprocal relationship between the f_{T} and both BV and V_{A}. Given a transistor design point where the base and emitter profiles are assumed constant, the f_{T} may be increased either by increasing the collector doping concentration or making the collector shorter (e.g., by decreasing the collector epilayer thickness), both of which delay the onset of the Kirk effect. Increasing the collector doping, decreases the Early voltage because of the increased base-width modulation. It also increases impact ionization, which lowers breakdown voltage. The reduction in collector epilayer thickness also increases the impact ionization due to the higher field from the same voltage supported over a shorter distance. For this reason, in one technology usually several devices differing only by the doping in the collector exist. In this manner, on the device level, speed can be directly traded for breakdown voltage. This tradeoff between f_{T} and BV is referred to as the "Johnson Limit" [24, 25] and states that, due to material limitations in carrier velocity and avalanche generation, the product of current gain cutoff frequency f_{T} and open base breakdown voltage $\mathrm{BV}_{\mathrm{CEO}}$ should be relatively constant.

Ongoing vertical profile scaling reduces carrier transit times and lateral scaling is reducing parasitics of the SiGe HBTs. Besides that, band gap engineering techniques in SiGe HBTs provide an extra degree of freedom in their device design [26]. Fundamentally, the strained SiGe layer reduces the base band gap, increases emitter injection efficiency, reduces emitter charge storage and reduces the base transit time. Therefore, for an identical emitter/base/collector dopant profile between Si BJT and SiGe HBT, the graded Ge profile in SiGe HBTs increases f_{T} degrading breakdown voltage with open base $\mathrm{BV}_{\mathrm{CEO}}$ only due to inevitable increase of the current gain. In contrast to this, increase of f_{T} by postponing the Kirk effect, also degrades the base-collector junction breakdown $\mathrm{BV}_{\mathrm{CBO}}$ due to increase of the built-in electric field caused by the enlargement of collector doping concentrations. Graded SiGe base transistors have the added advantage of higher V_{A} without compromising on f_{T} or BV. Further, power gain cutoff frequency f_{max} is as useful a figure of merit of bandwidth as f_{T} for wireless PA applications. SiGe HBT device physics does not demand a rigid tradeoff between breakdown voltage and f_{max}, implying significant device optimization opportunities [27] with respect to wireless PA applications. Thus, while BV and f_{T} provide a convenient metric to characterize SiGe HBT ruggedness and speed, they do not define strict device performance boundaries for PA applications.

Breakdown voltage dependence on bias and driving conditions

Since the speed of modern SiGe:C HBTs is increased (partially) on the expense of the breakdown voltage the collector-emitter bias voltage in recent designs often approaches or even crosses the open base collector-emitter breakdown $\mathrm{BV_{CEO}}$ point [28], which is of minor practical interest for modern analog circuit designers, because in many cases it must be exceeded to meet the target specifications. The bipolar device is fundamentally limited by the avalanche multiplication in the collector-base region [29] and therefore circuits should be designed to always unconditionally operate below base-collector breakdown voltage $\mathrm{BV_{CBO}}$. Designs above $\mathrm{BV_{CEO}}$ are, as is going to be seen, common practice in many wireless PA applications.

When avalanche multiplication takes place in the reverse biased base-collector junction (of npn type transistors), the generated holes drift toward the p-doped base and electrons toward n-doped collector region. Once in the base region, a hole may take either of two courses: it may exit the device through the base contact without further reaction or recombine with an electron and trigger additional injection of electrons from emitter into base, which is basically the current amplification action. The injected electrons will contribute to the increased avalanche multiplication and cause a further increase of electron injection from the emitter, forming a positive feedback loop. The strength of the positive feedback, which modulates the breakdown voltage, increases with increasing external impedance seen by the base electrode since the avalanche-generated holes are increasingly forced to stay withing the device with larger base terminal impedance. The same holds for pnp type transistors, with respective electron-hole substitution. Therefore, the configuration of base connection, which affects the base terminal impedance, has a direct impact on the breakdown voltage. Note, that this positive feedback in breakdown is a unique feature for bipolar transistors that is not found in FETs or diodes. The difference lies in the fact that the breakdown path in bipolar transistors traverses through two p-n junctions and, more importantly, one of the junctions usually remains forward-biased, enabling current amplification. Open-base configuration corresponds to infinite external impedance and maximized positive feedback, rendering the corresponding breakdown voltage $\mathrm{BV_{CEO}}$ to be the smallest breakdown voltage across collector and emitter. However, this configuration is rarely found in most practical circuit applications, and this explains why $\mathrm{BV_{CEO}}$ does not frequently serve as the voltage limit. The opposite extreme happens with the base shorted to emitter, where the external impedance is effectively zero and the positive feedback is absent, leading to the corresponding breakdown $\mathrm{BV_{CES}}$ largest and in ideal case (infinite intrinsic base conductance) equivalent to open emitter base-collector breakdown voltage $\mathrm{BV_{CBO}}$, but however in reality always somewhat lower than that. The most realistic case is the one of a configuration lying in between the two extreme cases. This is the one in which the base is connected to the emitter through a finite resistance/impedance value. The dependence of the bipolar transistor breakdown voltage on the source impedance can be exploited in practical PA design to significantly increase the safe operating voltage range.

When biased with zero impedance, for example the ideal voltage source between base and emitter nodes, the net base terminal current, due to avalanche can change the sign, that is to say, the flow direction. At the voltage bias point when in this case base current changes sign and equals zero the collector-emitter breakdown voltage at open base bias configuration occurs [30]. In other words the same mechanism is responsible for both of the effects and the asymptotic infinite collector current in the open base terminal case corresponds to the zero base terminal current in the short-cutted base-emitter case. Also if a bipolar transistor is driven by the current source in the emitter, the breakdown point is even more postponed. Key benefits of a cascode amplifier are derived from and implied by the last fact.

Although junction breakdown is the most common avalanche-originated mechanism that affects the safe operation region of devices, there is another phenomenon related to the avalanche that may also limit the operation of bipolar transistors. The phenomenon, often referred to as the pinch-in effect, refers to a situation in which the vertical current path in the intrinsic device is abruptly squeezed into the very center of the device when base-collector voltage exceeds a certain value [31]. When the device is placed in a common-base forced-emitter current configuration, direction of the net base electrode current is reversed when the avalanche-generated hole current in npn and electron current in pnp type devices, becomes larger than the current supplied from outside. As the reverse base current becomes sufficiently large and the device enters the deep avalanche region, emitter crowding effect takes place due to the lateral voltage drop across the finite base layer resistance. Since the current direction is reversed, the voltage drop occurs from the center to the edge of the active area, opposite of normal operation conditions, resulting in the current crowded at the center of the device instead of the edge of the emitter. This pinch-in mechanism tends to occur in an abrupt fashion and causes a sudden drop of collector current and base-emitter voltage, altering the bias condition of the device. Hence, the described pinch-in effect may potentially limit the voltage allowed across the emitter and collector for the forced-emitter configuration.

Role of junction electric field and impact ionization

Higher operating frequencies demand for faster devices (with higher unity gain frequencies) that, for transistors, as discussed, further imply lower breakdown voltages and therefore, lower power supply voltages. This is an issue for RF designers because of reduced signal-to-noise levels. Low breakdown voltage (together with noise considerations) of the device caused by high built-in junction electric fields influences the dynamic range of operation and is absolutely key issue for analog RF applications.

It is worth noting that operation at high electric fields is the backbone of modern semiconductor electronics. Indeed, (high) electric field accelerates minority carriers to saturation velocity and accordingly shortens their transit times through a space charge region where they are mainly moved by the drift, for which the base-collector junction is a typical example. Increasing output voltage of a BJT tends to widen the base-collector depletion layer and narrow the quasineutral base width where minority carriers are predominantly moved by diffusion [32], leading to increased transit time across the base-collector space charge region, but reduces transit times across the quasineutral base. It is accompanied by the increase of intrinsic base resistance and by the reduction of base-collector capacitance as well. Since the two effects tend to compensate each other, the combined effect of these trends are mostly balanced out, but still relatively small speed increase can be observed [33] in most of the devices.

Output voltage increase is accompanied by an enlargement of the electric field. If the junction electric field is raised above the value of the critical electric field, charge carriers are accelerated gaining energies enough for ionization to occur. A free carrier (an electron or a hole) impacts on the atom of a semiconductor. If the energy of the carrier is large enough, this carrier will knock out the electron from the valence shell of the atom. As a result, two newly formed free carriers, an electron and a hole, appear. In other words, if an initial carrier has enough energy, it can initiate the transition of an electron from the valence band to the conduction band. The minimal energy necessary to carry out the act of impact ionization is called threshold energy [34]. So, the electric field is the one responsible of transferring an energy to drift electrons, but it is the accumulated energy of an electron that is responsible for impact ionization. Hence, from the law of energy conservation follows that the threshold energy cannot be lower than the energy gap of a semiconductor material.

Diode versus transistor in terms of avalanche multiplication

Avalanche that occurs in isolated p-n junctions or diodes is not of the same level of complexity as in transistors, but it is of the same origin. For impact ionization to be initiated, free carriers that could be accelerated are necessary. If there are no such carriers, carrier multiplication cannot be initiated. While through the reversely biased (isolated) p-n junction the only current that can be multiplied is the minority carrier saturation current, through a transistor in a forward active regime of operation there is always a main current that can be multiplied. Consequently, since the current flowing through the junction that breaks down is to a large extent independent of its bias, there is one more degree of freedom and therefrom avalanche multiplication in transistors is, in a way, more complex than that in diodes. Therefore, the mechanism for a diode and a transistor is different for this type of breakdown which is not going to be the case with another type of breakdown, as will be discussed later in the text.

1.2.2 Tunneling breakdown

A general trend of doping concentration increase, observed in both pure CMOS and BiCMOS processes is very likely to continue [35]. In MOSFETs doping concentration in the channel is increased in order to combat short channel effects. In bipolar transistors, the vertical doping profiles of all three regions on the main current path are aggressively increased in order to increase transistor speed, however the highest doping concentrations remain in the emitter and base regions.

In a p-n junction where one or both sides are highly doped, under reverse bias an overlap of the energy band edges may occur, that is the valence band edge of the p-type material can have larger energy level than the n-type conduction band edge. In such scenario quantum-mechanical tunneling of electrons through the band gap spontaneously occurs. Valence band electrons may tunnel through the forbidden energy gap from p to n side preserving the energy level they had. In such a manner an electron (the tunneling one, in the conduction band) hole (previous place of the tunneled electron) pair is created adding up to the current flow. This process of carrier generation, in contrast to avalanche generation, does not depend on the current flowing through the space charge region. It only depends on the electric field, which is determined by the level of doping concentrations on p and n side and applied voltage between them.

Band-to-band tunneling (BtBT), as it occurs in reversely biased p-n junctions made from highly doped p- or n-type semiconductor material contributes to the total junction leakage current. Other identifiable contributions are the Shockly-Read-Hall (SRH) recombination and trap-assisted tunneling (TAT) at relatively lower biases and junction avalanche at relatively higher biases. The middle range between the two, is dominated by the band-to-band tunneling which has relatively weak temperature dependence, as opposed to SRH recombination and TAT, and thereby it is easily distinguishable on the current-voltage plot over multiple temperatures.

Leakage is very important in both MOSFETs and HBTs. In CMOS drain-to-body junction is naturally biased in reverse both in digital and (RF) analog applications. It has been suggested [36] that the junction leakage will present a fundamental limit in scaling of the traditional MOS transistor structure. On the other hand, in bipolar transistors, the higher doped base-emitter junction is in forward in the forward active regime. This is a usual mode of operation of BJTs in analog circuits. Nevertheless, there are exceptions from this rule in, for example class B or C power amplifiers (of C-BiCMOS technology) or power control circuits where for the part of the cycle the transistor is off and the base-emitter junction is reversely biased.

As noted, in some applications the base-emitter junction may be switched between forward and reverse bias. With reverse bias, a relatively high electric field is established laterally across the peripheral emitter-base junction. Once electron-hole pairs are generated either by thermal emission from traps or tunneling [37], the carriers are accelerated within the high field region and become hot. In a silicon oxide interface is located in proximity to that region, as is often the case for typical bipolar structures, the hot carriers generate traps by breaking weak interfacial bonds. The increased trap density enhances the carrier recombination rate and, as a result, base current is increased and current gain is reduced in the low base-emitter voltage bias region. Although the base leakage current does not result in significant change in most RF characteristics [33], it may degrade some parameters such as noise.

1.3 Semiconductor device modeling

Mathematical modeling, that uses mathematical language to describe certain system, is not limited to use only in the natural sciences and engineering disciplines (such as physics or electronics), also in the social sciences (such as economics, psychology or sociology) the use of mathematical models is extensive. It can be defined as a representation of the essential aspects of an existing system (or system to be constructed) which presents knowledge of that system in usable form.

Electronic design automation (EDA) is the category of tools for designing and producing electronic systems ranging from printed circuit boards (PCBs) to integrated circuits (ICs). It is sometimes referred to as computer-aided design (CAD). CAD can be roughly divided on TCAD (technology CAD) oriented toward manufacturability process flow and semiconductor device design and ECAD (electronic CAD) whose aim is more drawn to electrical circuit design. Backbone of both is the use of (physics-based) mathematical models for describing physical phenomena.

Technology computer-aided design is a branch of EDA that models semiconductor fabrication and semiconductor device operation. The modeling of fabrication is termed Process TCAD, while the modeling of the device operation is referred to as Device TCAD. The core of the process part of the TCAD is modeling of processing steps such as diffusion or ion implantation, that combined in a simulation of a process flow yields a (multidimensional) semiconductor structure (non seldom accompanied by metals and insulators). Physical mechanisms within the created structure with non homogeneous properties like dopant concentrations or material type can then be simulated as a function of applied bias conditions, temperature, light emission, etc. In simulation, solid-state physical equations such as (multidimensional) continuity or Poisson equation are solved to produce an insight in the device operation.

The goals of TCAD start from the physical description of an integrated circuit process flow and devices used in circuits, considering both the physical configuration and related material properties, and build the links between the broad range of physical and electrical behavior models that support circuit design. Physics-based modeling of devices in distributed form is an essential part of the IC process development, as it enables to speed-up, better understand and finally refine, usually extremely expensive and time consuming, process flow(chart). It seeks to quantify the underlying understanding of the technology and abstract that knowledge to the device design level. The TCAD models are aimed to be as accurate as possible due to their primary use in device design. Device simulation speed is then traded for accuracy directly in the process of mesh creation and refinement, nonetheless TCAD device simulation is intended for accurate simulations of a single device, or a coherent structure of multiple devices, rather than a complete system or a complicated circuit. The models that are used for integrated circuit design will be described in the next section.

1.3.1 Compact semiconductor device models

Electronic computer-aided design, on the contrary from technology computer-aided design, is focused on the higher level design paradigms, namely in simulating electrical circuit and system behavior. Circuit designers and system architects take advantage of circuit simulators like SPICE (Simulation Program with Integrated Circuit Emphasis) in a process of design and verification of integrated circuits. Their intention, in contrast to process designers, is not to design a particular device, rather a complete circuit that can consist of thousands of transistors. For semiconductor devices found in such circuits, a TCAD description would be too slow, even for a number of mesh nodes providing minimal accuracy. Therefore, compact semiconductor device models that predict behavior of a design are used. Compact device models are by the rule orders of magnitude faster than the TCAD physical device models. Analog circuit simulators such as SPICE use compact models in simulating circuits. Most of the design work is related to integrated circuit designs which have a very large tooling cost, primarily for the photomasks used to create the devices, and there is a large economic initiative to get the design working without any iterations. Complete and accurate models are precondition for designs to work after the first run.

Modern integrated circuits are usually very complex. The performance of such circuits is difficult to predict without accurate computer models, including but not limited to models of the semiconductor devices used. The semiconductor device compact models include effects of transistor layout, like width, length, interdigitation, proximity to other devices, and others. Using such compact models transient, harmonic balance, AC sweep or DC current-voltage characteristics may be simulated, effects of parasitic device capacitances, resistances, and inductances, on circuit design may be studied, as well as time delays and temperature effects, to name a few items.

In order to reduce development cost and time to market, modern industrial electronic design efforts rely on circuit simulations. Hence, exists the need for industrially supported, advanced compact bipolar transistor models capable of describing relevant characteristics of modern SiGe HBTs in the relevant regimes of operation.

Transistor compact models have to be accurate and computationally efficient, that is, simple. Clearly, there is always a tradeoff between (model) accuracy and simplicity. For this reason a hierarchy of models of different levels of accuracy/complexity could be offered to a circuit designer. In order to avoid any (non)convergence problems that may occur within a simulator, the mathematical equations representing the device compact model must be continuous, with desirably continuous derivatives up to highest order (smooth in a mathematical sense) which are required by the Newton-Raphson algorithm. Since the transistor sizes may differ as well, one model should be capable of fitting all device sizes used in the actual design practice.

The combined requirements of computational efficiency and available memory restrict the device models for circuit simulations [38] into physical, empirical and lookup table models. Practically all the models used in today's circuit simulators fall into physics-based analytical model [39] category and range from simple to more complex models. This is the type of model which will be developed and mostly used in this thesis. The advantages of physics-based models [40] are that they are continuous and scalable and can be themselves used in predictive way. In such models noise and statistical prediction is inherently present after valid model parameters are extracted. Scalability can come in terms of device geometry and operating temperature. Implemented noise modeling in well-constructed physics-based compact models works out of the box. By statistical prediction, for example consequences of statistical spread in emitter resistances can be predicted. Same holds for all parameters. This only works if parameters can be physically interpreted. Finally, physics-based compact models can be used in predictive way to virtually predict circuit behavior of future technologies. For example, TCAD device models of some future generation process are modeled by the compact models that are then used within the circuit simulator to get an insight in potential benefits on a circuit or a system level. On the other hand, the disadvantage is that they are technology dependent and takes considerable time to develop the model when technology goes through fundamental change or simply moves outside the model validity range. Furthermore, effects resulting from new device structures often require minor or major modification of the existing model and may even require a new model. Also, parameter extraction, that will be addressed in the next subsection, for such models often consist of subtle steps that cannot be automated. In an empirical model, the equations representing device characteristics are purely of the curve fitting type and are thus not based on device physics. In a table lookup model the device current data are stored for different bias points and device geometries in a tabular form, obviously as a drawback having enormous amount of memory and time usurped for a range of devices over geometry and temperature. The later two types of compact models are of seldom use (with microwave X-parameter nonlinear models that recently gained popularity as exceptions) in present day circuit simulators and as such are not going to be subject of further elaboration of this thesis.

Physics-based device compact models describe the terminal behavior of a device in terms of current-voltage ($I-V$) and capacitance-voltage ($C-V$) characteristics based on carrier transport processes which take place within the device. These models thus reflect device behavior in all regions of operation of the device. Due to two-dimensional (2-D) and three-dimensional (3-D) nature of the physical effects governing electrical behavior of modern transistors, it is very difficult to obtain a closed form analytical formulation which is valid in all operating regions of interest. However, one can still obtain closed form analytical models of separate physical effects, based on device physics, that are generally valid only over a limited region of device operation. Despite this limitation, such models are frequently used for circuit simulators because of ease of computation. In order to create a full description of a compact model [41], these models of separate physical phenomena are combined into equivalent circuit of the device. Equivalent circuits describe electrical properties of the device by connecting circuit elements in an organized manner such that the complete model emulates the electrical terminal behavior of the device. The elements of equivalent circuits are not necessarily derived from closed form analytical function describing physics but can be also using an empirical approach. Equivalent circuit compact models are often used in circuit simulators to represent device characteristics because of the ease of evaluation. SPICE exclusively uses models with equivalent circuit description. Equivalent circuit models of certain elements of semiconductor devices are highly nonlinear and can be strongly dependent on bias, frequency, geometry or temperature.

Physics-based compact models are usually conceived as large signal ones. Afterwards, utilizing process of linearization of such models, the simulator can perform small signal AC analysis which produces a linear response of the simulated circuit. The large signal compact model can be directly evaluated to obtain DC bias solution or DC sweep over multiple biases or parameter values. To such large signal model, any analysis, say harmonic balance, can be applied in order to produce nonlinear large signal steady state solutions, or time dependent transient response, from the transient analysis. The most comprehensive compact models are those that describe large signal transistor behavior because all other simulation types follow just as special cases.

1.3.2 Model parameter extraction and optimization

Various terms such as device characterization, parameter extraction, optimization and model fitting are used to address tasks of preparing a given compact model to be actually employed in circuit simulation. In all these terms, the starting point would be a mathematical model that describes a certain semiconductor device. Such model has a number of parameters which are varied or adjusted to match the terminal characteristics of, for example, a particular transistor or set of transistors. The act of determining an appropriate set of model parameters is what is called parameter extraction. Afterward, the model, with its particular set of model parameters representing a particular transistor, is used in a circuit simulator to predict how a circuit with a certain kind of transistors at a given bias conditions will behave.

The precondition for parameter extraction or their optimization are corresponding measurements from which parameters are estimated or to which model parameters are fitted. One characterizes a device and in such a way obtains measurements necessary for estimating parameter values. In general, for the extraction of a full parameter set of a modern physics-based transistor compact model, C-V, DC bias I-V and S-parameter measurements are needed. To determine the parameters of geometry or temperature scaling rules, part of the measurements has to be repeated at at least one other (then reference) geometry or temperature, respectively. Sometimes, additional test structures are available for estimation of certain parameters for which no straightforward extraction strategy exists. In applications where compact models are used to predict future technology circuit behavior, instead of measurements, TCAD process and device simulations are used in producing generic data on which parameters can be extracted or optimized. Once the measured (or simulated) data are available, a parameter extraction or optimization strategy is used to find the best set of model parameter values to fit the data.

The terms extraction and optimization are often used interchangeably in the semiconductor industry, however, strictly speaking, they do not have exactly the same meaning [42]. By optimization is usually meant using generalized least-squares curve fitting methods such as the Levenberg-Marquardt algorithm [43, 44] to find a set of model parameters. By extraction, on the other hand, any technique that does not use general least-squares fitting methods is considered. This is a somewhat loose interpretation of the terms extraction and estimation. The main point is that there exist approximate equations that when (again in general, approximately) solved allow to get the extracted results in closed form.

Extraction has the advantage of being much faster than optimization, but it is not always as accurate. Since with modern computers optimization over thousands of measured and calculated/simulated points is in the order of seconds or minutes, rather than hours, the optimization drawback is far from being crucial. With optimization if anything goes wrong one can always change the range of data, weighting, upper and lower bounds, etc. which is not the case with extraction. More experienced users will usually prefer the flexibility, control, and accuracy that optimization provides, rather than go for one click extraction solution, that is less accurate.

Commercial software is available that provides both extraction and optimization together in the box. The idea here is to first use extraction techniques to make reasonable initial guesses and then use these results as a starting point for optimization, because optimization can give very poor results if poor initial guesses for the parameters are used. Nothing is wrong with using extraction techniques to provide initial guesses for optimization, but for an experienced user this is rarely necessary, assuming that the nonlinear least-squares curve fitting routine is robust (converges well) and the experienced user has some knowledge of the process under characterization. In addition, subsets of parameters should be obtained in the proper order so that those obtained at later steps do not affect those obtained at earlier steps. Also, optimization should never be used to obtain a parameter value when the parameter can be measured directly, as for example MOS oxide thickness.

Accuracy of a device model in predicting device characteristics is very much dependent on the accuracy of the model parameter values being used. Complexity of the models used in circuit simulators have increased significantly in the past decades. Further, most compact models used in circuit design are semiempirical analytical models containing various fitting parameters that do not have a well-defined physical meaning, and the number of these fitting parameters increases with the complexity of the model. Very often some of these fitting parameters become redundant, and no unique value can be determined for those parameters. Therefore, care must be taken in extracting or optimizing model parameter values from device data so that physical meaning of the parameter is retained as much as possible.

In general, for sophisticated models determining model parameters is an extremely important task on which the quality of the model and its ability to predict work conditions accurately is highly depended. Hence, special attention should be paid to parameter extraction for newly created models.

1.4 Motivation for breakdown modeling

The motivation for this thesis directly derives from the previous two sections. Models giving accurate prediction at the edge of breakdown become increasingly interesting for circuit designers. In that way they can increase confidence in their high frequency, high power analog or highly scaled (and hence high leakage power) digital designs.

Breakdown studies

In general, studies of breakdown can be grouped in three classes, according to the regime in which the device is biased and by the signal type that is driving it. The first class would then be the static, direct current (DC) class, in which in fact no time-dependent signal is applied, that is well-covered by the semiconductor literature. In the other classes, on the other hand, a time dependent signal is applied. The second class could be referred to as small signal alternating current (AC), where the device is biased in the breakdown regime and a small alternating signal is applied. In the third class the device is biased outside the avalanche regime but the applied AC signal is large enough to put the device into the breakdown regime during signal swings or vice versa so that device cannot be linearized as in the previous class.

Avalanche breakdown modeling

Due to extra noise source in cases where impact ionization multiplication is present, transistors are always kept biased outside the avalanche region in design of low noise amplifiers (LNAs). In contrast to LNAs, where noise levels are absolutely essential, in the design of power amplifiers (PAs) the focus is more drawn on the power itself. As seen from above considerations, because of the larger breakdown voltage of BJTs over MOSFETs, but as well as larger active gain [45] and lower noise, HBTs are the devices of preference for (mmW) RF integrated circuit designs.

For the mentioned reasons, in LNA design only an indication of breakdown would suffice. On the other hand in design of PAs which requests high output power or high output voltage swing, accurate modeling of breakdown is essential. The bipolar transistor Gummel-Poon model [46] did not address avalanche modeling (and heterojunction features found in modern HBT devices) in its initial release, nevertheless it still remains in wide use today due to its integration within SPICE. However, for more sophisticated (RF) designs, more advanced compact models like VBIC [47, 48] (Vertical Bipolar InterCompany) model, HiCuM [49] (High Current Model), or Mextram [50] (Most EXquisit TRAnsistor Model) have to be used. HiCuM and Mextram are to date world standard compact models of (heterojunction) bipolar transistors chosen by the Compact Model Council [51, 52] (CMC).

Avalanche modeling in all three of the models is similar (but not the same), differing between one another rather slightly than fundamentally. Mainly for numerical stability motivated by avoiding divergence issues, the models are restricted to the weak avalanche case [53, 54], the case where carriers generated in a process of impact ionization do not generate extra carriers. The largest difference between Mextram and HiCuM/VBIC models of avalanche current is that later [55] do not take into account the cases where at higher current densities the maximum value of electric field occurs at the buried layer [56] rather than at the base-collector junction. Also within the extended avalanche modeling of Mextram the decrease of the effective epilayer width due to base-widening is accounted for. All three of the models are of a planar breakdown occurring in the internal transistor, that is below the emitter, which is a reasonable assumption because published measurements for (self-aligned) poly-silicon emitter transistors show such a planar breakdown rather than a breakdown at the periphery of the external base-collector junction.

All three avalanche current representations of the mentioned compact models are using a conventional impact ionization local electric field modeling approach which is based on the presumption that carriers are instantaneously energized to the steady-state kinetic energy corresponding to the local electric field intensity. In a spatially or time varying electric fields, however, the carrier energy lags the field because of the finite energy relaxation time [57]. Spatially, this is the case in highly scaled BJTs where very high electric fields and field gradients exist [58] in the collector's epilayer region. Employing nonlocal impact ionization modeling is of fundamental practical use in TCAD device simulations. In compact models, since the model parameters are intended to be fitted to the measured curves, spatial nonlocal effects can be (somewhat) absorbed in effective values of the extracted parameters, which nonetheless often give (equally) good fit in the region they are designed to do so.

For the driving condition with fixed emitter current, or in the more general case of a high source impedance found at the emitter electrode (which holds for several often used practical transistor stages like common base cascode stages, emitter followers, transadmittance stages with strong negative feedback and to a certain extent emitter-coupled differential stages), the breakdown voltage at a given emitter current is higher than for a constant base-emitter voltage drive condition, but not nearly as high as expected from the classical theory [59]. This is because at a critical base current the emitter current abruptly pinches in to a small area in the center of the emitter. This effect is, as is the breakdown in the case of constant base-emitter voltage, caused by a lateral voltage drop across the base resistance but can no longer be modeled by a simple two-dimensional transistor model. Instead, a three-dimensional (3-D) model is required [60]. The onset of lateral instabilities is well defined by the critical base current which can be calculated by analytical relations [31]. This has been verified by measurements as well as by quasidistributed three-dimensional (3-D) transistor model (QDTM) consisting of a lot of intrinsic transistor elements. Being more suitable because of its efficiency, a six transistor model (6TM) [60] can be used to approximate QDTM. HiCuM's Level4 model version inherently models distributed effects through a proprietary sectionalized model. Also, simplification of the QDTM while preserving accuracy, but lowering computational complexity must be possible.

Going from simple direct current (DC) considerations, that are nonetheless highly relevant in the alternating signal domain, toward alternating current (AC), many other important physical mechanisms emerge as important or even decisive ones for analog integrated circuit design. Although transistors are becoming more and more sophisticated, subtle parasitic and distribution effects often have a profound role in the frequencies of interest for modern RF designs. In the small signal AC regime, distribution of the collector resistance over the base-collector capacitance can significantly impact transistor two-port parameter characteristics [61] in general and the modeling of the cutoff frequency and power gain in particular.

The influence of avalanche on small signal AC transistor characteristics has not been explicitly covered by the semiconductor literature. However, it can be expected that the modeling of distributed capacitances over resistances in the collector region, if it already makes an impact on transistor two-port parameter characteristics, would also impact the part of the small signal AC transistor characteristics where avalanche is dominant. Influence of avalanche on large signal AC transistor characteristics, besides ones which analyze the effect of avalanche on power amplifier nonlinearities [62, 63], has not been extensively addressed by the technical literature. The accurate large signal model verification could actually only be done on an advanced measurement setups, like (active) load-pull [64], where input and output impedances are precisely known. Since such measurements are nontrivial they are based on heavily exploited simulations beforehand. Performing such simulations is necessary to adjust all the power levels and generally gain an impression what can be expected from the experiment and which knowledge can be gained from it. An advanced compact model would be of interest for performing such simulations. The large signal behavior of the model could be also verified on such high-end measurement setup.

Tunneling breakdown modeling

Given the fact that, due to various reasons, leakage power became a significant portion of the total power in highly scaled digital systems, it is imperative for digital circuit designers and system architects to possess accurate prediction of the system's leakage mechanisms to continue to reap the benefits of technology exponential downscaling [65] as long as possible. On the other hand, since the digital chips nowadays consist of several billion transistors, models have to be increasingly efficient as well and have to be easily skipped if not evaluated without compromising on solver's convergence. The main leakage mechanisms in MOSFETs, gate, junction and sub-threshold leakage, will continue to increase as transistors are scaled down within technology nodes towards 10 nm [66]. The junction leakage in today's generations are mainly present due to the pocket implants, also called (super-)halos, that are used to combat short-channel effects [67] and off-state leakage, while it is also suggested that this very leakage will present the fundamental limit for scaling of the traditional MOS transistor structure [36]. This leakage type, of reversely biased p-n junctions, so natural for drain-to-body junction [68], is implied mainly by Shockley-Read-Hall (SRH) generation / recombination and trap-assisted tunneling (TAT) at relatively low voltages and higher temperatures, by band-to-band tunneling (BtBT) in the middle voltage range, and finally by impact-ionization avalanche multiplication current at high voltages. As for digital design transistors are almost never pushed even close to the avalanche breakdown point, accurate modeling of band-to-band tunneling leakage current is a must in order to correctly predict circuit behavior. Since tunneling current, due to hot charge carriers found in it, can damage silicon oxide, creating traps in it, and hence it may lead to a device degradation in terms of nonreproducibility, studies of band-to-band tunneling have to be carried out on carefully selected devices.

Junction parameter extraction

Both breakdown types, avalanche and tunneling, are driven by high electric field values in the space charge region of a p-n junction where that particular breakdown also occurs. Barely every single present day physics-based compact model of electric field is based on p-n junction depletion capacitance parameters. Therefrom, the accuracy of the extracted depletion capacitance parameters will have a great impact on the electric field model and consequently on the breakdown model accuracy as well. Hence, special attention should be paid to extraction of these parameters.

Junction parameters are usually optimized to the measured depletion capacitance value using nonlinear regression techniques. Temperature scaling parameter, the band gap is on the other hand usually extracted from measurements of ideal forward bias p-n junction current temperature dependence. However, various strategies to do so exist and obtained parameter accuracy greatly depends on the employed estimation methodology. Having a need to extract also temperature or geometry scaling rule, it is unclear whether to estimate the parameter governing this dependency simultaneously. Choosing the best strategy is not trivial so it is quite important to know which of the extraction methodologies are favorable to use from the statistical point of view.

1.5 Thesis outline

The presented work focuses on modeling of breakdown phenomena in semiconductor devices. In particular it discusses frequency limitations of local impact ionization modeling (in silicon). It also provides a reduction technique for avalanche breakdown as it may occur in multidimensional bipolar junction transistor structures. Then, it explains characterization, modeling and repercussions of working within the avalanche regime of bipolar transistors and applying an alternating signal. Further, contribution has been made to modeling of arbitrary p-n junction tunneling breakdown. Parameter extraction techniques, especially those incorporated in the junction electric field model, extensively used in modeling of both breakdown types, are analyzed and useful suggestions are provided. The material is organized as follows.

Chapter 2 summarizes a physical basis of impact ionization modeling, bringing focus on both hydrodynamic energy-transport semiconductor equations and drift-diffusion model which is the preferred one in compact modeling. Frequency limitations of local impact ionization models are examined and estimated. Modeling of avalanche breakdown in terms of ionization rates is presented through a compact model perspective, what provides a basis on which subsequent two chapters advance.

Chapter 3 concentrates on a quasidistributed bipolar transistor model reduction technique whose aim is to greatly reduce computational cost while preserving the original model accuracy. Such models are usually used to model devices in which current crowding effect, that can lead to a vertical current pinch-in, may occur.

Chapter 4 elaborates on bipolar transistors pushed into an impact ionization regime and found in an alternating signal environment. Specifically, in such cases avalanche characterization is important in order to proceed with further analysis of any kind that concentrates on it. Addressed are the necessities for accurate modeling of such regimes. Repercussions of avalanche on some important intrinsic active device properties from circuit design prospective are explained in depth.

Chapter 5 contains the description of a novel model for the band-to-band tunneling current in a p-n junction. It consists of the model physical foundations, model implementation and finally its verification on state of the art industrial and modern in-house devices. The model is fully physics-based, it is smooth in a mathematical sense on a whole real domain, features increased efficiency without compromising accuracy, and innovative parametrization which greatly improves scaling over geometry and temperature.

Chapter 6 is devoted to the analysis of parameter extraction strategies. Since this work concentrates on modeling of breakdown phenomena that are both driven by the electric field within the p-n junction's depletion region, accent is drawn to the p-n junction parameters and their extraction methodologies. The estimation strategies are compared in statistical terms which provide an insight how the two, or more, can be assessed and compared, and which one would be more suitable for use in practice.

Chapter 7 finally collects the main conclusions of the thesis and provides the reader with several recommendations for future work.

Chapter 2

Physical basis of impact ionization modeling

MENTIONED is already in the previous introductory chapter that in fact two different paradigms of semiconductor device modeling can be identified. They differ in the level of abstraction and in the end target. One level would be simulating the electrical characteristics of devices, as response to external electric, thermal or optical boundary conditions which are imposed on the structure. This is done by solving semiconductor (differential/integral) equations between the each pair of nodes in the generated mesh. The other level would be composed of compact models that intend to describe the device behavior based on given terminal conditions in closed analytic form expressions. This chapter introduces modeling of impact ionization effects in both of these paradigms. Its purpose is to create an overview of used techniques and comment on them, as well as to prepare the basis for two subsequent chapters.

Electron-hole pair production caused by impact ionization requires a certain threshold field strength and the possibility of acceleration, that is, wide space charge regions. If the width of a space charge region is greater than the mean free path between two ionizing impacts, charge multiplication occurs, which can further cause electrical avalanche breakdown. The reciprocal value of the mean free path is referred to as ionization coefficient. Impact ionization generation rate can be expressed using these coefficients. Afterward, such impact ionization generation rates can be used both within a semiconductor TCAD model or a compact one.

2.1 Impact ionization in semiconductor modeling

It is of great importance to predict limitations of certain models' physical validity. Impact ionization models are not exception to this. The main goal of this section is to reach an analytical expression for temporal limitation of the local impact ionization model and analyze it. The other goal is to make the reader familiar to the places within semiconductor transport models where impact ionization may occur.

2.1.1 Semiconductor transport equations

The hierarchy of semiconductor transport models usually start from Boltzmann transport equation as the most fundamental one that is solved by Monte Carlo simulations, going toward more simple and more specific forms from hydrodynamic over thermodynamic to drift-diffusion models that can be analytically solved. The common part in the mentioned transport models are the Poisson's and continuity equations.

Poisson and continuity equations

The three governing equations for charge transport in semiconductor devices are the Poisson equation and the electron and hole continuity equations. The differential form of the Poisson's equation for electrostatic potential φ and electric field $\mathbf{E}$ is written as

$$\Delta\varphi = \nabla \cdot \nabla\varphi = \nabla^2\varphi = -\nabla \cdot \mathbf{E} = -\frac{q}{\varepsilon}\left(p - n + N_\mathrm{A} - N_\mathrm{D}\right) - \frac{\rho_S}{\varepsilon} \quad , \tag{2.1}$$

where ε is the electrical permittivity, q is the elementary charge, n and p are the electron and hole densities, N_D is the concentration of ionized donors, N_A is the concentration of ionized acceptors and ρ_S is the charge density contributed by traps and fixed charges. The electron and hole continuity equations are expressed as

$$\begin{aligned} \nabla \cdot \mathbf{J}_n &= q\left(R + \frac{\partial n}{\partial t}\right) = q\left(R' - G_n^{\mathrm{II}} + \frac{\partial n}{\partial t}\right) \quad , & (2.2)\\ -\nabla \cdot \mathbf{J}_p &= q\left(R + \frac{\partial p}{\partial t}\right) = q\left(R' - G_p^{\mathrm{II}} + \frac{\partial p}{\partial t}\right) \quad , & (2.3) \end{aligned}$$

in which $\mathbf{J}_n$ and $\mathbf{J}_p$ are electron and hole current densities, G_n^{II} and G_p^{II} are the impact ionization generation rates for electrons and holes, respectively, R is the net electron-hole recombination rate and R' is the electron-hole recombination rate excluding impact ionization. Current density is the movement of charge density. The continuity equation says that if charge is moving out of a differential volume, that is divergence of current density is positive, then the amount of charge within that volume is going to decrease, so the rate of change of charge density is negative. Therefrom, the continuity equations amount to a conservation of electric charge.

Drift-diffusion semiconductor transport model

The drift-diffusion (DD) model is widely used for simulation of carrier transport in semiconductors and is defined by the basic semiconductor equations (Poisson and continuity ones), where the current densities for electrons and holes are given by

$$\begin{aligned} \mathbf{J}_n &= -qn\mu_n\nabla\phi_n = q\left(n\mu_n\mathbf{E}_n + D_n\nabla n\right) \quad , & (2.4)\\ \mathbf{J}_p &= -qp\mu_p\nabla\phi_p = q\left(p\mu_p\mathbf{E}_p - D_p\nabla p\right) \quad , & (2.5) \end{aligned}$$

where μ_n and μ_p are the electron and hole mobilities, D_n and D_p are the electron and hole diffusivities, and ϕ_n and ϕ_p are the electron and hole quasi-Fermi potentials, respectively. Neglecting effects of the band gap narrowing and assuming Boltzmann carrier statistics $\mathbf{E}_n = \mathbf{E}_p = \mathbf{E} = -\nabla\varphi$ by Helmholtz decomposition.

Thermodynamic semiconductor transport model

A first step toward the complete hydrodynamic energy transport model would be the thermodynamic (TD) or nonisothermal model that extends the drift-diffusion approach to account for electrochemical effects under the assumption that the charge carriers are in thermal equilibrium with the lattice. Therefore, the carrier temperatures and the lattice temperatures are described by a single temperature T_L. In this model besides the basic set of partial differential equations (PDEs), the lattice heat flow equation is employed as well. The drift-diffusion set of equations for current densities are generalized to include the temperature gradient as a driving term

$$\begin{aligned} \mathbf{J}_n &= -qn\mu_n \left(\nabla\phi_n + P_n \nabla T_L \right) \quad , & (2.6) \\ \mathbf{J}_p &= -qp\mu_p \left(\nabla\phi_p + P_p \nabla T_L \right) \quad , & (2.7) \end{aligned}$$

where P_n and P_p are absolute thermoelectric powers (also called thermopowers) for electrons and holes, respectively.

Hydrodynamic semiconductor transport model

With continued downscaling into the deep submicron range, neither internal nor external characteristics of certain state of the art semiconductor devices can be described properly using the conventional drift-diffusion transport model. In particular, the drift-diffusion approach cannot reproduce velocity overshoot and often overestimates the impact ionization generation rates. The Monte Carlo methods for the solution of the Boltzmann kinetic equation are the most general approach, but because of their high computational requirements, they cannot be used for the routine simulation of devices in an industrial setting. In this case, the hydrodynamic (or energy balance) model provides a very good compromise. Among many variations of this model there is the full formulation which includes the so-called convective terms and consists of eight PDEs, while the simpler form without the convective terms includes only six PDEs. In the hydrodynamic model, the carrier temperatures T_n and T_p are assumed not to equal the lattice temperature T_L. Equations for electron and hole current densities are updated accordingly to account for contributions due to the spatial variations of electrostatic potential, electron affinity, the band gap, as well as to take into account contributions due to concentration gradient, the carrier temperature gradients and the spatial variation of the effective masses. Together with basic semiconductor equations three additional equations can be solved to find the temperatures. In general, the model consists of basic set of PDEs and the energy conservation equations for electrons, holes and the lattice. The energy balance equations read

$$\begin{aligned} \nabla \cdot \mathbf{S}_n &= \mathbf{J}_n \cdot \nabla E_C - H_n - n\frac{w_n - w_0}{\tau_{wn}} - \frac{\partial (n w_n)}{\partial t} \quad , & (2.8) \\ \nabla \cdot \mathbf{S}_p &= \mathbf{J}_p \cdot \nabla E_V - H_p - p\frac{w_p - w_0}{\tau_{wp}} - \frac{\partial (p w_p)}{\partial t} \quad , & (2.9) \\ \nabla \cdot \mathbf{S}_L &= H_L + \frac{w_n - w_0}{\tau_{wn}} + \frac{w_p - w_0}{\tau_{wp}} - \frac{\partial (c_L T)}{\partial t} \quad , & (2.10) \end{aligned}$$

where $\mathbf{S}_n$, $\mathbf{S}_p$ and $\mathbf{S}_L$ are energy fluxes, that is energy flow densities, for electrons, holes and the crystal lattice, respectively.

$$\mathbf{S}_n = -\frac{5}{2}\left(\frac{k_{\mathrm{B}}T_n}{q}\mathbf{J}_n + \frac{k^2}{q}n\mu_n T_n \nabla T_n\right) \quad , \tag{2.11}$$

$$\mathbf{S}_p = -\frac{5}{2}\left(\frac{k_{\mathrm{B}}T_p}{q}\mathbf{J}_p + \frac{k^2}{q}p\mu_p T_p \nabla T_p\right) \quad , \tag{2.12}$$

$$\mathbf{S}_L = -\kappa_L \nabla T_L \quad , \tag{2.13}$$

where κ_L is the thermal conductivity of the respective material. Electron and hole energy relaxation times are given in terms of τ_{wn} and τ_{wp} which can be assumed to be constant in the first and most common approximation. The lattice specific heat c_L is also non seldom approximated by a constant value. Average energy densities of electrons, holes and carriers that are in equilibrium with crystal lattice are respectively defined as

$$w_n = \frac{3}{2}k_{\mathrm{B}}T_n \quad , \quad w_p = \frac{3}{2}k_{\mathrm{B}}T_p \quad \text{and} \quad w_0 = \frac{3}{2}k_{\mathrm{B}}T_L \quad , \tag{2.14}$$

where k_{B} is the Boltzmann constant. Electron, hole and lattice temperatures are denoted as T_n, T_p and T_L, respectively. Energy gain/loss terms due to generation/recombination processes are denoted by H_n, H_p and H_L. Usually these effects are small, especially H_L which is very often neglected and hence will not be subject of discussion. The terms for electrons and holes are usually represented as

$$H_n = Rw_n \text{ or } H_n = R'w_n + E_g G_n^{\mathrm{II}} \quad \text{and} \quad H_p = Rw_p \text{ or } H_p = R'w_p + E_g G_p^{\mathrm{II}} \, , \tag{2.15}$$

where E_g is the effective band gap energy. As in current continuity equations, as can be seen, in energy balance model equations effects of impact ionization are essentially introduced by the terms G_n^{II} and G_p^{II}, representing the net electron and hole generation rate per unit volume. However, it seems that there is no clear consensus on what would be the most appropriate functional dependence of H_n and H_p terms on impact ionization generation rates, only that H_L is independent of it.

Strictly speaking, the energy balance equations (2.8), (2.9) and (2.10), together with the corresponding Poisson's (2.1) and current continuity equations (2.2) and (2.3), represents an example of the energy-transport (ET) hydrodynamic (HD) semiconductor model. Such a model is obtained when in the course of deriving the momentum equation from Boltzmann transport equation the average kinetic energy of the carriers is neglected against their thermal energy and the average total carrier energy can be expressed as in (2.14). Within the ET HD semiconductor model, the impact ionization term G^{II} not only affects the spatial and temporal distribution of the carrier concentrations, as in DD transport model, but it also perturbates their corresponding average energies. Indirectly, the spatial and temporal distribution of the electric field and lattice temperature are also affected by impact ionization generation rates.

To further increase the efficiency of the semiconductor simulators using ET model some additional pragmatical simplifications have been proposed. They will be demonstrated in this place for the electron energy flux continuity equation. First, heterojunctions are excluded and band gap narrowing is neglected, thereby gradient of conduction and valence bands became identical $\nabla E_C = \nabla E_V = \mathbf{E}$ and equal to electric field. Neglecting also the thermal flux, second term between the brackets in equations (2.11) and (2.12), as well as the recombination and generation terms R and H_n and H_p, the energy balance equation for electrons becomes

$$-\frac{5}{2}\frac{k_\mathrm{B}}{q}\left(qT_n\frac{\partial n}{\partial t}+\mathbf{J}_n\cdot\nabla T_n\right)=\mathbf{J}_n\cdot\mathbf{E}-\frac{3k_\mathrm{B}}{2}\left[n\frac{T_n-T_L}{\tau_{wn}}+\frac{\partial(nT_n)}{\partial t}\right] \quad . \tag{2.16}$$

The equation for holes looks likewise. Expressing also the electron (and hole) current density in terms of the average electron (and hole) velocity as

$$\mathbf{J}_n=-qn\mathbf{v}_n \quad \text{and} \quad \mathbf{J}_p=qp\mathbf{v}_p \quad , \tag{2.17}$$

the electron energy balance identity (2.16) can be further rewritten as

$$n\frac{\partial T_n}{\partial t}-\frac{2}{3}T_n\frac{\partial n}{\partial t}+\frac{5}{3}n\mathbf{v}_n\cdot\nabla T_n+\frac{2q}{3k_B}n\mathbf{v}_n\cdot\mathbf{E}+n\frac{T_n-T_L}{\tau_{wn}}=0 \quad . \tag{2.18}$$

The first two terms in the last equation are directly responsible for the temperature dynamic behavior. In most practical situations it is reasonable to assume that

$$2T_n\frac{\partial n}{\partial t}\ll 3n\frac{\partial T_n}{\partial t} \quad \text{and} \quad 3p\frac{\partial T_p}{\partial t}\gg 2T_p\frac{\partial p}{\partial t} \quad , \tag{2.19}$$

yielding the energy balance equation for electrons [57] in the form

$$\frac{\partial T_n}{\partial t}+\frac{5}{3}\mathbf{v}_n\cdot\nabla T_n+\frac{2q}{3k_B}\mathbf{v}_n\cdot\mathbf{E}+\frac{T_n-T_L}{\tau_{wn}}=0 \quad , \tag{2.20}$$

having no direct dependence on the electron concentration n. Assuming that the average electron velocity is instantaneously defined by the local value of the electric field, $\mathbf{v}_n=\mathbf{v}_n(\mathbf{E})$ (the momentum relaxation time is much smaller than the energy relaxation time), the last equation governs the spatial and temporal dependence of the electron temperature solely as a function of the electric field.

The simplified energy balance equation (2.20) is particularly useful in the postprocessing methodology for impact ionization modeling. The idea is to first evaluate the electric field distribution solving a reduced set of semiconductor equations and than to determine the carrier temperature in the postprocessing procedure solving expression (2.20), thereby reducing the total model evaluation time.

2.1.2 Temporal nonlocal temperature response

The spatial dislocation of the carrier temperature and electric field distribution is well-known and mostly analyzed in the stationary case. Here, a simple analysis of the temporal nonlocal response is presented. Let us consider, as a special case, a homogeneous semiconductor multiplication region, having uniform electron temperature distribution ($\nabla T_n=\nabla T_p=0$), subject to a (rapidly) time-varying electric field. The energy balance partial differential equation for electrons (2.8) is further simplified to a constant-coefficient first order linear ordinary differential equation (ODE)

$$\frac{d\left[T_n\left(t\right)-T_L\right]}{dt}+\frac{T_n\left(t\right)-T_L}{\tau_{wn}}=h_n(t) \quad , \tag{2.21}$$

where the time dependent function on the right-hand side is given by

$$h_n\left(t\right) = -\frac{2q}{3k_\mathrm{B}}\mathbf{v}_n(\mathbf{E}\left(t\right))\cdot\mathbf{E}\left(t\right) \quad . \tag{2.22}$$

The solution of nonhomogeneous first order linear ordinary differential equation with constant coefficients (2.21) is given by the following integral

$$T_n(t) = T_L + \int\limits_0^t h_n(\zeta)\,\exp\left(\frac{\zeta - t}{\tau_{wn}}\right)\,d\zeta \quad . \tag{2.23}$$

For the simple sinusoid electric field excitation with angular frequency of $\omega = 2\pi f$, assuming linear dependence of electron velocity on the electric field

$$h_n(t) = h_{n0}\cos\left(\omega t\right) \quad , \tag{2.24}$$

the corresponding temporal response of the electron temperature is

$$T_n(t) = T_L + \frac{T_{n0}}{\sqrt{1+\omega^2\tau_{wn}^2}}\cos\left(\omega t - \psi\right) \tag{2.25}$$

where

$$T_{n0} = h_{n0}\tau_{wn} \tag{2.26}$$

is the electron temperature at $\omega = 0$ and

$$\psi = \arccos\left(\frac{1}{\sqrt{1+\omega^2\tau_{wn}^2}}\right) \tag{2.27}$$

is the phase delay of the electron temperature harmonic response.

Notice from (2.25) that the magnitude of the electron temperature harmonic response decreases with frequency. Figure 2.1 shows the normalized transfer characteristics of the electron temperature frequency response for several constant values of energy relaxation time, typically found[69] in semiconductor literature.

Solution (2.25) can be used to define a frequency of the electron temperature after which temporal nonquasistatic effects become nonnegligible. This frequency can be defined as, say, the frequency at which the response magnitude is half of that in the low frequency limit (−3 dB point on the graph). At sufficiently high operational frequencies it is in principle possible to suppress the heating of carriers and correspondingly increase the, so-called, dynamic breakdown voltage. However, for present state of the art SiGe HBT devices it still lies in the region, if not beyond the cutoff frequency then at least, outside operational and measurement range. Hence, based on this analysis there is no evidence to believe that the current avalanche compact models that do not include nonlocal temporal effects are not suitable for use in design of advanced RF analog and mixed-signal circuits.

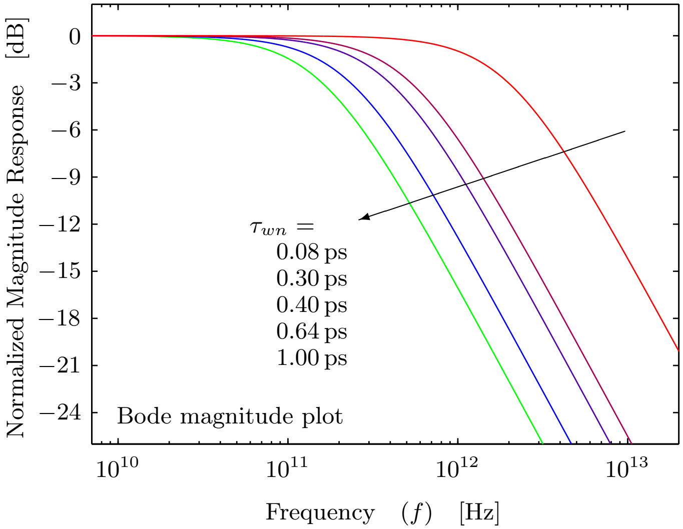

Figure 2.1 – Bode plot of normalized temperature magnitude (frequency) response to the sinusoid excitation for five different values of electron energy relaxation time typically found in semiconductor literature [69]. Depending on the actual energy relaxation time value, the nonlocal effects become dominant above different frequency points.

2.2 Description of impact ionization phenomena

In general, there exist three main and at once mostly used approaches for studying impact ionization mechanisms and phenomena in semiconductor devices

- Approximation of the characteristic breakdown electric field
- Monte Carlo simulations
- Impact ionization rates approximation

which are going to be explained further in this section.

Approximation of the characteristic breakdown field

Approximation of the characteristic breakdown field is the simplest way of describing impact ionization mechanism and avalanche breakdown phenomenon. In the framework of this approach one assumes that avalanche breakdown occurs if the maximum electric field at any point in the semiconductor structure exceeds the value known as the characteristic breakdown field. On the other hand, it is assumed that is no impact ionization exists nor occurs at all if the maximum field is lower than the characteristic breakdown field.

Even this simplest approximation can provide explanation [34] of a tendency that the characteristic breakdown field and thereby breakdown voltage increases with temperature for most semiconductors and semiconductor structures. The breakdown voltage increases monotonically with temperature, so that the smaller the doping level, the stronger is the temperature dependence of it.

The characteristics of the dependencies can be explained by simple qualitative considerations. As the impact ionization process is defined by the energy of the carrier, gained from the electric field between scattering collisions, the probability of impact ionization decreases as the scattering events become more frequent. Thus, since the frequency of phonon scattering increases with temperature, it becomes more difficult for an electron (hole) to take a large amount of energy from the electric field. This can be described formally as a decrease in carrier energy relaxation time as temperature increases. As a result, the breakdown field and breakdown voltage increase with temperature. The lower the doping level, the larger is the relative contribution of phonon scattering to the total scattering process. That is why the temperature dependence of the breakdown voltage becomes greater as the doping level decreases. There are several relatively rare exceptions to this rule which can be very dangerous from the point of view of possible thermal instabilities.

Monte Carlo transport simulations

Although very powerful numerical method that allows to simulate any transport phenomena in semiconductors, including impact ionization and breakdown effects it is rarely used to calculate operating regimes of devices due to extremely high complexity and hence huge consumption of the processor time. It is usually used to check the principal problems and to calculate ionization rates. It has also been successfully used to simulate ultra small nanoscale semiconductor devices when all other techniques have failed due to large space inhomogeneities and very high space derivatives that are characteristic for such devices. It is becoming an optional attribute of many commercial TCAD device simulators. The accuracy of its calculations is constrained only by the understanding knowledge of band structure and scattering rates. This type of simulation is, whatsoever, out of scope of the present thesis.

2.2.1 Impact ionization rates approximation

The approximation of impact ionization rates practically presents an effective compromise between the oversimplified approach of the effective breakdown field and rigorous but rather complicated Monte Carlo simulation procedure. In the framework of this approach one assumes that impact ionization characterized by ionization rates of α_i for electrons and β_i for holes, which are defined as probabilities of impact ionization per unit length. In the local modeling approach, ionization rates are assumed to be instant functions of the electric field. This assumption has obvious limitations.

If the electric field is instantly increased from zero to a certain value, it takes some finite time for an electron/hole to acquire the threshold energy which is necessary to produce an elementary act of impact ionization. Roughly speaking, this time will be equal to the energy relaxation time which is in order of picoseconds in high electric fields. Therefrom, when considering processes on frequencies of hundreds of gigahertz and higher, local impact ionization models may not be sufficient.

A similar situation emerges if an electric field E changes very sharply in space. It is clear that if electric field changes notably along the carrier mean free path it will be impossible to say which value of the field should be used to calculate $\alpha_i(E)$ or $\beta_i(E)$. Taking a characteristic mean free path of $\lambda \approx 10^{-8}\,\mathrm{m}$ and a characteristic electric field $E_i \approx 10^7\,\mathrm{V/m}$, one can estimate a characteristic magnitude for dE/dx of about $10^{15}\,\mathrm{V/m^2}$. Such large values of dE/dx are realized either in extremely small semiconductor structures with characteristic sizes of around 10 nm or in rapidly varying applied bias conditions. In these cases Monte Carlo simulation should be used to describe the ionization processes correctly.

The approximation of impact ionization rates is nevertheless the most popular and efficient tool in studying impact ionization phenomena in its region of applicability.

Even in very strong electric fields it is as a rule the case that only a small portion of the electrons/holes possess energy which exceeds the characteristic critical energy W_0 necessary for impact ionization to occur. On average the carrier energy is much smaller, and it is limited by optical phonon scattering.

It is often suggested that approximation of ionization rates are empirical, while in fact this is only partially true. Namely, starting from the concept of statistical physics known as the mean free path, one can arrive to the expressions for ionization rates, what will be presented in further detail.

The mean free path λ may be defined as the average distance traveled by a particle (in case of semiconductors electrons or holes) between two consecutive collisions with other particles (or lattice). It can be assumed that collisions occur randomly, so that a particle has the same chance of collision in any interval of length. The average number of collisions per unit length is $1/\lambda$. The probability that a collision occurs in an interval dl would therefore be $p(dl) = dl/\lambda$, whereas the probability that no collisions occur is, of course $q(dl) = 1 - dl/\lambda$. Now if $q(l)$ is the probability that no collisions occur in l, then for the whole interval $l + dl$ it may be written:

$$q(l+dl) = q(l)q(dl) = (1 - dl/\lambda)\, q(l) \quad . \tag{2.28}$$

Identity $q(l + dl) = q(l)q(dl)$ is justified because the events of collision in a length l and collision in length dl are independent, that is conditional probability is just the product of two probabilities. Now expanding the left-hand side of previous equation in a Taylor series and neglecting terms after the linear one, the expression is simplified to a first order autonomous (and hence also linear) ordinary differential equation

$$\frac{dq(l)}{dl} = -\frac{1}{\lambda} q(l) \quad , \tag{2.29}$$

which has solution in form of a Poisson probability distribution

$$q(l) = \exp\left(-\frac{l}{\lambda}\right) \quad \Longleftrightarrow \quad p(l) = 1 - \exp\left(-\frac{l}{\lambda}\right) \quad . \tag{2.30}$$

Hence, for the sake of control, the average distance traveled between collisions is an integral over all distances that a particle traveled a certain distance without collision and then collided in the next infinitesimal distance. It is expressed in probabilistic terms as

$$\langle l \rangle = \int_0^\infty l\, dp(l) = \int_0^\infty \frac{l}{\lambda} q(l)\, dl = \int_0^\infty \frac{l}{\lambda} \exp\left(-\frac{l}{\lambda}\right) dl = \lambda \quad , \tag{2.31}$$

which merely confirms the coherence of analytical calculations.

In order to achieve an energy value of W_0, an electron (or a hole) driven by electric field has to travel without collision for a distance

$$l = W_0/qE \tag{2.32}$$

Having in mind the previous derivation, the probability of such an event is

$$q(l) = \exp\left(-\frac{l}{\lambda}\right) = \exp\left(-\frac{W_0}{q\lambda E}\right) = \exp\left(-\frac{E_0}{E}\right) \quad , \tag{2.33}$$

with $E_0 = W_0/q\lambda$ being the characteristic value of the electric field. Hence, the expressions for ionization rates for electrons and holes take forms

$$\alpha_i(E) = \alpha_\infty \exp\left(-\frac{E_{n0}}{E}\right) \quad \text{and} \quad \beta_i(E) = \beta_\infty \exp\left(-\frac{E_{p0}}{E}\right) \quad . \tag{2.34}$$

The experimental $\alpha_i(E)$ and $\beta_i(E)$ dependencies for the most important semiconductor materials are usually described by the following empirical equations:

$$\alpha_i(E) = \alpha_\infty \exp\left[-\left(\frac{E_{n0}}{E}\right)^{m_n}\right] \quad \text{and} \quad \beta_i(E) = \beta_\infty \exp\left[-\left(\frac{E_{p0}}{E}\right)^{m_p}\right] \quad . \tag{2.35}$$

In silicon for example $m_n = m_p = 1$ [70]. The last two expressions present the so-called Chynoweth's law [71] for the ionization coefficients.

It is worth noting that if electric field E is relatively small $E \ll E_0$, ionization rates α_i and β_i will be very strongly dependent on the actual strength of the field, while for $E \approx E_0$ they will show fairly weak dependence on it. In very strong fields $E \gg E_0$, on the other hand, α_i and β_i tend towards their limiting values of α_∞ and β_∞, respectively. These limiting values correspond to a situation in which the distance between two elementary acts of impact ionization $1/\alpha_\infty$ or $1/\beta_\infty$ is close or equal to the mean free path λ, the case where electrons/holes ionize at every scattering act.

2.2.2 Multidimensional ionization rate approximations

Approximation of multidimensional ionization rates are based on above derivations, done for a one-dimensional case. They also divide into local and nonlocal approaches.

Local modeling

In the local modeling approach is it assumed that the carrier generation terms are directly defined by the local current densities and electric field distribution as

$$G^{\mathrm{II}} = G_n^{\mathrm{II}} + G_p^{\mathrm{II}} \quad \text{where} \quad G_n^{\mathrm{II}} = \alpha_i\left(E\right)\frac{|\mathbf{J}_n|}{q} \quad \text{and} \quad G_p^{\mathrm{II}} = \beta_i\left(E\right)\frac{|\mathbf{J}_p|}{q} \quad . \tag{2.36}$$

The electric field dependence of the ionization coefficients $\alpha_i(E)$ and $\beta_i(E)$ is similarly expressed by the multidimensional Chynoweth's law

$$\alpha_i = \alpha_\infty \exp\left[-\left(\frac{E_{n0}\,|\mathbf{J}_n|}{\mathbf{E}\cdot\mathbf{J}_n}\right)^{m_n}\right] \quad \text{and} \quad \beta_i = \beta_\infty \exp\left[-\left(\frac{E_{p0}\,|\mathbf{J}_p|}{\mathbf{E}\cdot\mathbf{J}_p}\right)^{m_p}\right] \quad , \tag{2.37}$$

with α_∞, β_∞, m_n, m_p, and E_{n0}, E_{p0} as model parameters. As may be noted from previous equations, the driving force in multidimensional impact ionization can be computed as the component of the electrostatic field in the direction of the current (scalar dot product of the two vectors).

The local modeling approach can be used with the simple drift-diffusion carrier transport equations without energy balance equations. It has been also employed as a basis for the development of the existing transistor compact models. However, the assumption that the ionization coefficients depending only on the local electric field strength is applicable only if the electric field $\mathbf{E}$ varies slowly in space and time. The localized and highly varying electric fields require the nonlocal modeling approach.

Nonlocal modeling

The carrier generation caused by impact ionization is fundamentally described by carrier energy (temperature) distribution and not by the local electric field intensity. There are several expressions (one of them is Schöll-Quade model [72]) that relate the net carrier generation rate to the carrier temperature. Apart from the local modeling approach, the net impact ionization generation rate is controlled in the nonlocal modeling approach by the carrier temperature and concentration instead of the local electric field and carrier current density. The model requires employment of the full set of the ET HD semiconductor transport equations.

The original nonlocal models often introduce complex expressions which are computationally costly. Therefore, they are usually not appropriate for use within the compact semiconductor device models. Thus, some simplified versions are employed non seldom, including there empirical nonlocal impact ionization modeling.

Empirical nonlocal modeling

The idea of the empirical nonlocal modeling approach is to transform the empirical expression (2.37), from the local modeling approach, into the analogous expression in terms of carrier energy (temperature). Namely, in the stationary and homogeneous case, and if the generation/recombination terms R' and H_n and H_p are neglected, the energy balance equation for electrons (2.8) becomes

$$\mathbf{E}\cdot\mathbf{J}_n = n\frac{w_n - w_0}{\tau_{wn}} = \frac{3}{2}k_{\mathrm{B}}n\frac{T_n - T_L}{\tau_{wn}} \quad . \tag{2.38}$$

On the other hand, the intensity of the electron and hole current density in terms of the average intensity of electron and hole velocity $v_n = |\mathbf{v}_n|$ and $v_p = |\mathbf{v}_p|$ is

$$|\mathbf{J}_n| = qnv_n \quad \text{and} \quad |\mathbf{J}_p| = qnv_p \quad , \tag{2.39}$$

From equations (2.38) and (2.39) the multidimensional Chinoweth's formula for the net electron impact ionization generation rate becomes

$$\alpha_i = \alpha_\infty \exp\left[-\left(\frac{w_{n0}}{w_n - w_0}\right)^{m_n}\right] = \alpha_\infty \exp\left[-\left(\frac{T_{n0}}{T_n - T_L}\right)^{m_n}\right] \quad , \tag{2.40}$$

where

$$w_{n0} = \frac{3}{2}k_{\mathrm{B}}T_{n0} = q\tau_{wn}v_nE_{n0} \quad , \tag{2.41}$$

is the impact ionization critical electron energy (temperature). Analogous expression can be derived also for the hole impact ionization generation rate β_i.

Although the ionization coefficient (2.40) has been derived using local relationship between the electric field and electron temperature (2.38), the model is practically employed using the temperature distribution governed by the full energy balance equation. The physics-based nonlocal modeling approach should be always a preferred choice for the detailed device modeling. Nevertheless, the empirical nonlocal modeling approach seems to provide the easier path for eventual implementation of nonlocal effects into compact models.

2.3 Compact impact ionization modeling

The present, state of the art, compact modeling of impact ionization and avalanche breakdown phenomena is restricted to one-dimensional (1-D) multiplication regions along p-n junction's depletion region (or bipolar junction transistor's epilayer).

2.3.1 Multiplication factor

Let us consider fluxes of electrons and holes that pass through a certain region of a semiconductor. While traveling a distance dx, each electron will create an average of $\alpha_i dx$ electron-hole pairs. The increase in the electron current density j_n due to electron multiplication and analogously the same current increase due to hole multiplication will be

$$\left.\frac{dJ_n}{dx}\right|_n = \alpha_i J_n \quad \text{and} \quad \left.\frac{dJ_n}{dx}\right|_p = \beta_i J_p \quad . \tag{2.42}$$

Hence it can be written,

$$\frac{dJ_n}{dx} = \alpha_i J_n + \beta_i J_p \quad , \quad \text{and analogously} \quad \frac{dJ_p}{dx} = -\alpha_i J_n - \beta_i J_p \quad , \tag{2.43}$$

from which subsequently follows the identity relating current spatial derivatives

$$\frac{dJ_n}{dx} = -\frac{dJ_p}{dx} \quad , \tag{2.44}$$

where of course the total current density is just the sum of the two $J = J_n + J_p$.

Simultaneously solving the set of the last three equations and the Poisson's equation with appropriate boundary conditions will allow us to describe the steady state one-dimensional electron and hole distributions under avalanche multiplication.

With appropriate boundary and initial conditions the transient characteristics can be described by a set of first order partial differential equations

$$\begin{aligned} \frac{\partial j_n}{\partial x} &= q\left(R + \frac{\partial n}{\partial t}\right) = q\left(R' + \frac{\partial n}{\partial t}\right) - \alpha_n j_n - \alpha_p j_p \quad , & (2.45) \\ \frac{\partial j_p}{\partial x} &= -q\left(R + \frac{\partial p}{\partial t}\right) = -q\left(R' + \frac{\partial p}{\partial t}\right) + \alpha_n j_n + \alpha_p j_p \quad . & (2.46) \end{aligned}$$

Last two equations present the one-dimensional case of more general multidimensional continuity equations for electron (2.2) and hole (2.3) current densities, respectively.

Neglecting transient terms resulting from effects other than impact ionization avalanche multiplication and thereby going back to the stationary case one can write

$$\frac{dJ_n}{dx} = (\alpha_i - \beta_i) J_n + \beta_i J \quad \text{and} \quad \frac{dJ_p}{dx} = (\beta_i - \alpha_i) J_p + \alpha_i J \quad , \tag{2.47}$$

which is a general form of linear nonhomogeneous ordinary differential equation of order one with variable coefficients[†]. Its solution that yields current densities at arbitrary point in space is not as important for compact modeling as the ratio between electron/hole current density injected into certain region and the current density flowing out of that certain region. Multiplication factors for electrons and holes are defined as

$$M_n = \frac{J_n(W_\mathrm{D})}{J_n(0)} \quad \text{and} \quad M_n = \frac{J_p(0)}{J_p(W_\mathrm{D})} \quad , \tag{2.48}$$

where W_D is the width of the avalanche multiplication region. It is assumed that electrons are injected from the left while holes are flowing in from the right side. Boundary conditions are used to eliminate the constant of integration. In solving continuity equation for electrons (2.47) boundary condition $J_n(0) = J$, while boundary condition $J_p(W_\mathrm{D}) = J$ is used in solving hole static continuity equation (2.47). After integration, the obtained result can be expressed in an integral form

$$X_n = 1 - \frac{1}{M_n} = \int\limits_0^{W_\mathrm{D}} \alpha_i \exp\left[-\int\limits_0^{x} (\alpha_i - \beta_i)\, dz\right] dx \quad , \tag{2.49}$$

[†] Equation (2.47) has the form $y'(x) + P(x)y(x) = Q(x)$. The general solution is given in a form

$$y(x) = \left[\int Q(x)\left(\exp\int P(\varsigma)d\varsigma\right)dx + \kappa\right] \Big/ \exp\int P(x)\,dx \quad ,$$

which can also be used as the more general solution for equation (2.21) of the previous section.

$$X_p = 1 - \frac{1}{M_p} = \int\limits_0^{W_\mathrm{D}} \beta_i \exp\left[- \int\limits_x^{W_\mathrm{D}} (\beta_i - \alpha_i)\, dz \right] dx \quad , \tag{2.50}$$

while for equal electron and hole ionization rates $\alpha_i = \beta_i = \gamma$ it holds

$$1 - \frac{1}{M_n} = 1 - \frac{1}{M_p} = 1 - \frac{1}{M} = \int\limits_0^{W_\mathrm{D}} \gamma(x) dx \quad , \tag{2.51}$$

The avalanche breakdown voltage is defined as the voltage where multiplication factor M_n or M_p approaches infinity. Therefrom, the avalanche breakdown condition is given by the ionization integrals (2.49), (2.50) and (2.51) which should equal unity. Equations (2.49) and (2.50) are equivalent, meaning, the breakdown condition depends only on what is happening within the multiplication region and not on the carriers (or primary current) that initiate(s) the avalanche process. The situation is the same if the primary current of only one type or mixed one initiates the breakdown, so either expression (2.49) and (2.50) gives the breakdown condition.

For a homogeneous field distribution along the avalanche multiplication region and $\alpha_i = \beta_i = \gamma$ case, the breakdown condition takes the simplest form of $\gamma W_\mathrm{D} = 1$. This means that in this special case avalanche breakdown occurs when the electron/hole creates just one electron-hole pair on average while traveling through the avalanche region of length W_D. Namely, say, an electron creates a hole and the newly emergent hole in turn creates an electron, and so on. This positive feedback provides the appearance of an avalanche breakdown. It can be concluded in general that if the multiplication factor is approaching infinity it is not necessary to have any external carrier to support the avalanche breakdown, hence it is a self-supporting process.

2.3.2 Avalanche generated current

In the static or quasistatic case, when generation/recombination terms are neglected, the total current density flowing through a multiplication region can be expressed as

$$J = J_n(W_\mathrm{D}) + J_p(0) = M_n J_n(0) + M_p J_p(W_\mathrm{D}) \quad , \tag{2.52}$$

and from there only the avalanche current density may be written as

$$J_\mathrm{AVL} = (M_n - 1)\, J_n(0) + (M_p - 1)\, J_p(W_\mathrm{D}) \quad . \tag{2.53}$$

Assuming that the current is dominantly defined by a single carrier type (typically electrons, but sometimes holes as well), the avalanche current becomes

$$J_\mathrm{AVL} \approx (M_n - 1)\, J_n(0) \quad \text{or} \quad J_\mathrm{AVL} \approx (M_p - 1)\, J_p(W_\mathrm{D}) \quad . \tag{2.54}$$

For weak avalanche, when secondary impact ionization can be neglected and M_n or M_p are close to unity holds (in general written for multiplication factor M)

$$M \approx 1 \iff M\,(M-1) \approx M - 1 \iff M - 1 \approx 1 - \frac{1}{M} \quad , \tag{2.55}$$

and the avalanche current density if further simplified to one the following expressions

$$J_{\mathrm{AVL}} \approx X_n J_n(0) \quad \text{or} \quad J_{\mathrm{AVL}} \approx X_p J_p(W_{\mathrm{D}}) \quad , \tag{2.56}$$

almost exclusively used in present day compact modeling of avalanche multiplication.

2.3.3 Impact ionization integral

The most simple empirical models define the ionization integrals X_n and X_p in the depletion region of a p-n junction, in general, and in the base-collector junction of bipolar transistors, in particular, as

$$X_n = X_p = X = (V/\mathrm{BV})^N \quad , \tag{2.57}$$

where V is the applied junction voltage, BV is the breakdown voltage (for base-collector junctions often annotated as $\mathrm{BV_{CBO}}$, that is, the base-collector junction breakdown voltage at open emitter contact) and N is a parameter determined from the fit to experimental results. The expression (2.57) is not very accurate for silicon junctions especially for large values of multiplication factor. Moreover, the physical connection to spatio-temporal distribution of the electric field and charges is lost.

The alternative physics-based modeling approaches are based of the original ionization integral expression (2.49) and (2.50) and ionization coefficient in Chynoweth form (2.35). However, several additional assumption are introduced: (i) the model parameters m_n and m_p are close to unity, (ii) the vectors of the electric field and current density in (2.37) are collinear and (iii) the ionization coefficients for electrons and holes are equal. With this assumptions the ionization integrals are expressed as

$$X_n = \alpha_\infty \int\limits_0^{W_{\mathrm{D}}} \exp\left(-\frac{E_{n0}}{|E(x)|}\right) dx \quad \text{and} \quad X_p = \beta_\infty \int\limits_0^{W_{\mathrm{D}}} \exp\left(-\frac{E_{p0}}{|E(x)|}\right) dx \quad , \tag{2.58}$$

with the idea to evaluate the integrals on the right-hand side for certain approximations of the electric field distribution in the region of interest for impact ionization.

Assuming uniform doping in the epilayer side of bipolar transistor's base-collector junction and carrier transport with constant saturation electron velocity the absolute value of electric field has linear spatial dependence

$$|E(x)| = E_{\max}\left(1 - \frac{x}{\chi}\right) \quad , \tag{2.59}$$

where $E_{\max}$ is the maximum intensity of the electric field while $-1/\chi$ defines the slope of its linear dependence as shown in Figure A.1.

In order to analytically evaluate the integrals in (2.58), linear electric field distribution is approximated in the neighborhood of its maximum intensity $E_{\max}$ as

$$|E(x)| \approx \frac{E_{\max}}{1 + x/\chi} \quad , \tag{2.60}$$

and therefrom the ionization integral for electrons is obtained [53, 54] as

$$X_n = \frac{\alpha_\infty}{E_{n0}} \chi E_{\max} \left\{ \exp\left(-\frac{E_{n0}}{E_{\max}}\right) - \exp\left[-\frac{E_{n0}}{E_{\max}}\left(1+\frac{W_{\mathrm{D}}}{\chi}\right)\right]\right\} \quad . \tag{2.61}$$

Control physical quantities $E_{\max}$ and χ are possible to obtain from the solution of the Poisson's equation in bipolar transistor's epilayer region of the collector. Depending on the current density level, the maximum intensity of the electric field can be located [56] either at the base-collector junction or in the transition region between the epilayer and highly doped collector buried layer. The problem of determining spatio-temporal distribution of the electric field intensity in the epilayer is strongly related to the modeling of the junction depletion capacitances as well as modeling of the quasisaturation, that is, current dependency of the maximum intensity of the electric field $E_{\max}$ and the base pushout effect. How the necessary quantities are calculated in both cases of electric field maximum is explained in Appendix A in detail.

The compact modeling of the electric field distribution along the epilayer has been considered so far only by stationary or quasistatic analysis and a full dynamic analysis is still missing. Moreover, for high-speed bipolar devices with a very thin and highly doped epilayer the avalanche multiplication based on the local electric field distribution will be overestimated. Than this model will require nonlocal modeling corrections.

2.3.4 Nonlocal postprocessing of carrier temperatures

The simplified energy balance equation for electrons (2.20), this time assuming stationary or quasistationary case, in one dimension simplifies to

$$\frac{d\left[T_n(x) - T_L\right]}{dx} + \frac{T_n(x) - T_L}{\lambda_{wn}} = -\frac{2q}{5k_{\mathrm{B}}} E(x) \quad , \tag{2.62}$$

where $\lambda_{wn} = 5 v_n \tau_{wn}/3$ is the energy relaxation length that is in connection to the energy relaxation time. The solution of the upper differential equation is given as

$$T_n(x) = T_L - \frac{2}{5}\frac{q}{k_{\mathrm{B}}} \int\limits_0^x E(\zeta) \exp\left(\frac{\zeta - x}{\lambda_{wn}}\right) d\zeta \quad , \tag{2.63}$$

and it can be employed [58] in the postprocessing phase to evaluate the spatial dependence of the electron temperature for the given distribution of the electric field in combination with the empirical nonlocal modeling of the ionization coefficients.

2.3.5 Applications of avalanche current compact models

In general, described already existing compact model impact ionization solutions, that are as discussed in this section limited to weak avalanche description, are used in practically every modern physics-based compact model of p-n junction or bipolar/field effect transistor to describe avalanche generated current. Such avalanche/impact ionization compact model will be used as the basis for further development of models in two distinct directions which will be shown in the following two chapters.

Chapter 3

Distributed avalanche modeling in bipolar transistors

Reduction technique for an accurate modeling of complex effects manifested in an avalanche regime of bipolar transistors is shown in this chapter. Bilinear approximation is utilized to significantly reduce computational cost of the model made for precise consideration of breakdown phenomena. The simplification method is practically implemented on the basis of a vertical bipolar compact model Mextram. Extraction of additional parameters is studied. The reduction technique is quantitatively compared to the model from which it is derived and the results are presented.

3.1 Avalanche-induced instabilities

Modeling of breakdown phenomena is becoming a central problem in today's design of high-speed bipolar circuits. It is especially important for the output stages that should simultaneously provide ever increasing speeds and relatively high output signal power/voltage swing. The interplay of the device maximum operating frequency and output power requires complex design tradeoffs. To this end, accurate modeling is essential to fully exploit the potential of advanced Si and SiGe bipolar technologies, and to allow safe circuit design with bipolar transistors operating above the collector-emitter breakdown voltage at open base, $\mathrm{BV}_{\mathrm{CEO}}$.

Carrier impact ionization could change the direction and significantly increase the intensity of a transistor base current leading also to instabilities in the device behavior [59]. The main source of the device instabilities is the current crowding effect caused by a considerable lateral base current in the intrinsic transistor region. Because of the finite base resistance, it creates a nonuniform biasing along the base-collector junction. This effect is best visualized by performing three-dimensional (3-D) simulation of a typical bipolar transistor as shown in Figure 3.1, where different tones refer to different current densities in a vertical current flow.

Designers already possess a profound tool that has the possibility to give a re-

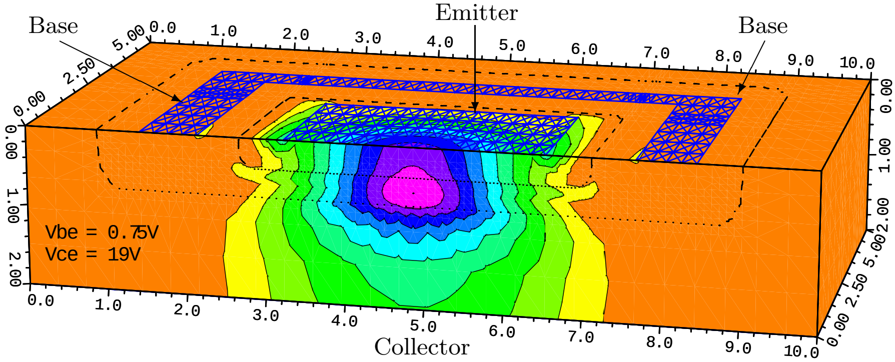

Figure 3.1 – Three-dimensional (3-D) technology computer-aided design (TCAD) simulation of a vertical bipolar junction transistor in the avalanche breakdown regime. Different grayscale shades (intensities) correspond to different current densities within transistor. The current density is larger in the center part underneath the emitter. This effect is causing the current pinch-in under emitter current driving configuration.

spectable prediction of device behavior in the avalanche operating regime thanks to the quasidistributed three-dimensional transistor model [31]. It precisely predicts the onset of avalanche-induced instabilities under any driving condition. Unfortunately, although very precise, this model suffers from extremely high complexity and inefficiency. This chapter shall give a contribution in efficient modeling of multidimensional avalanche effects. The main benefit shall be a model that is not too expensive for circuit simulation but nevertheless preserves the accuracy of previous models developed for the same purpose. The two models are compared in terms of simplicity and accuracy and verified by measurements on sophisticated test devices.

3.2 Distributed multitransistor model

One way to address multidimensional avalanche effects in circuit design is to employ sectionalized bipolar transistor models. The basic idea is to partition the transistor base under the emitter into vertical sections associating each with a separate intrinsic transistor model, as is shown in Figure 3.2. The bases of the neighboring sections are coupled to each other with an effective variable base resistance being a fraction of the total base resistance. The network of intrinsic transistors is capable of capturing the distributed character of the main transistor current, but its major drawback is its complexity. A circuit representation of the sectionalized intrinsic transistor model is shown in Figure 3.3, while the extrinsic part remains unmodified.

There have been attempts to the decrease quasidistributed bipolar transistor model complexity using problem symmetry and sacrificing the model accuracy using a rather small number of sections [60]. On the other hand, a novel technique to significantly reduce the model computational complexity/cost without compromising much with the accuracy of sectionalized transistor models is proposed in this chapter.

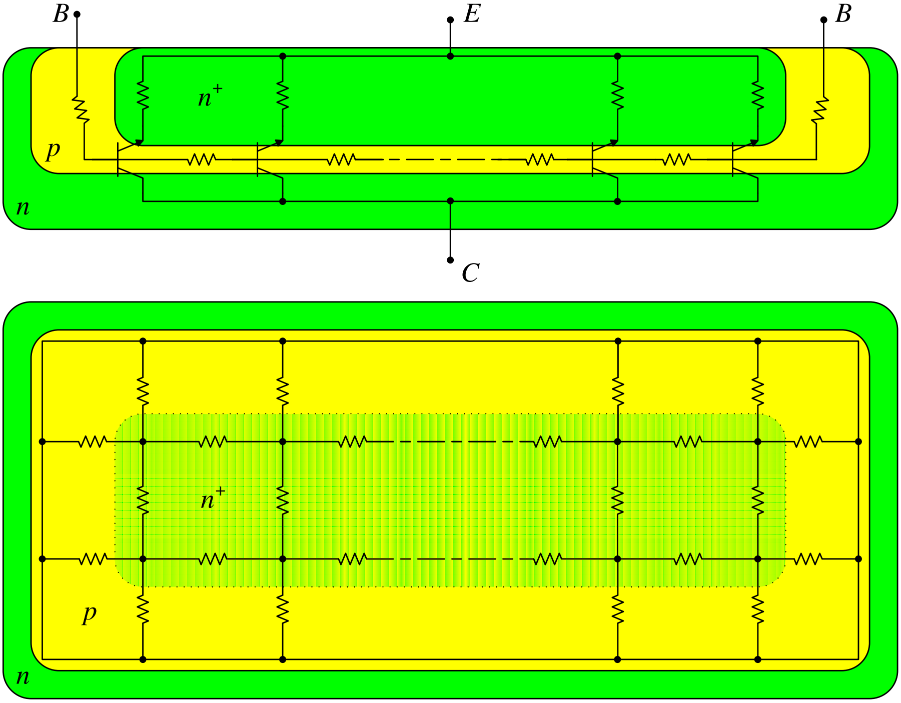

Figure 3.2 – Simplified cross-section side view (upper figure) and top view (lower figure) of a typical rectangular emitter shape bipolar junction transistor. The schematic of a sectionalized intrinsic transistor models together with a resistor network, both needed for an accurate prediction of the distributed avalanche effects, is drawn.

3.2.1 Bilinear approximation as a reduction technique

Firstly, if the emitter possesses a rotational symmetry of order s, that attribute should be utilized to reduce the multitransistor model complexity by a factor of s. In practice, since the emitter is often rectangular with an even number of base stripes (this will be assumed onwards if not indicated otherwise) the computational time shall be divided by a factor of four. Furthermore, a proposed method of reduction employs only two one-dimensional chains of full intrinsic transistor sections along the symmetry lines of the emitter contact using bilinear interpolation to get the effect of a full transistor matrix. Moreover, only four full intrinsic transistor elements (center one, corner one, one on the horizontal and one on the vertical axis corresponding to the adequate row and column of the element being interpolated) have been used to interpolate the particular inner transistor current. This is schematically depicted in Figure 3.4. Currents (base, emitter and collector) of the inner transistor segment in the row i and the column j, $i_x = i(i,\, j)$ are interpolated using the following expression

$$i_x = \frac{(i_h - i_f)\,(i_v - i_f)}{i_l - i_f} + i_f \quad , \tag{3.1}$$

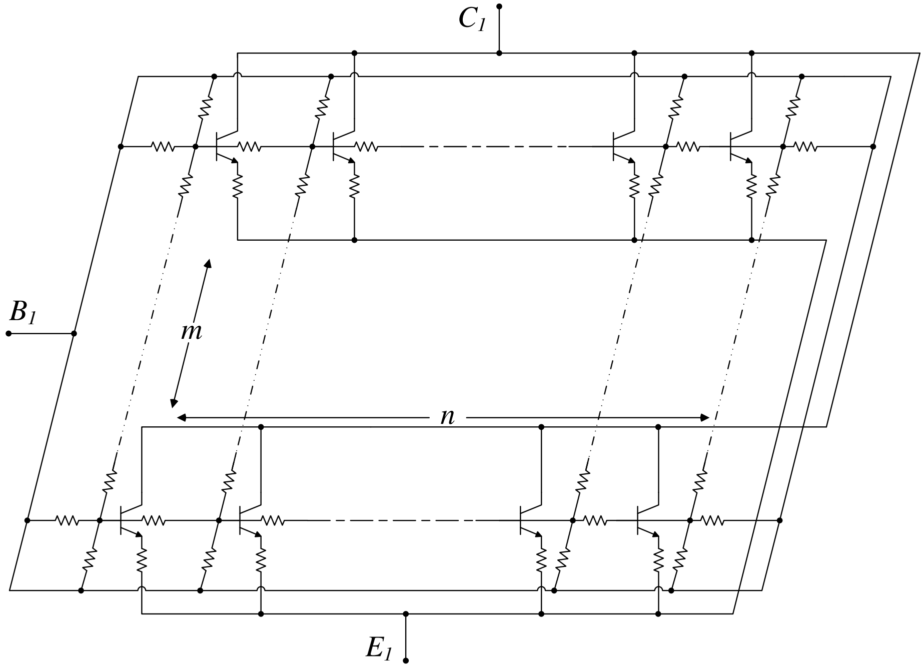

Figure 3.3 – The sectionalized intrinsic quasidistributed bipolar transistor model.

in which $i_f = i(1, 1)$ is the corresponding current of the corner full intrinsic transistor element, $i_f = i(m/2, n/2)$ current of the center intrinsic transistor element, $i_h = i(m/2, j)$ current of the corresponding horizontal transistor element, $i_v = i(i, n/2)$ current of the corresponding vertical element and m and n the number of rows and columns of the transistor matrix, respectively. Because the intrinsic transistors model does not have a substrate description, only two currents, for example base and emitter currents, are approximated with the third one, collector current, obtained by a simple addition operation of the two. It can be concluded that in the original segmented model the number of full intrinsic transistor elements would be $m \cdot n$, exploiting symmetry $(mn)/4$, and using proposed reduction method only approximately $(m + n)/2$ with $((m-1)(n-1)-1)/4$ interpolation segments.

3.2.2 Verilog-AMS implementation

The original and reduced segmented models have been practically implemented in an industry standard modeling language for analog circuits, Verilog-A(MS). Mextram, a physics-based vertical bipolar transistor model was used in the segmentation process, although HiCuM [49], VBIC [48] or some other compact model could be used as well. The intrinsic part of Mextram, highlighted in Figure 3.5, is used to create the full intrinsic transistor element model that is later multiplied. The extrinsic part of

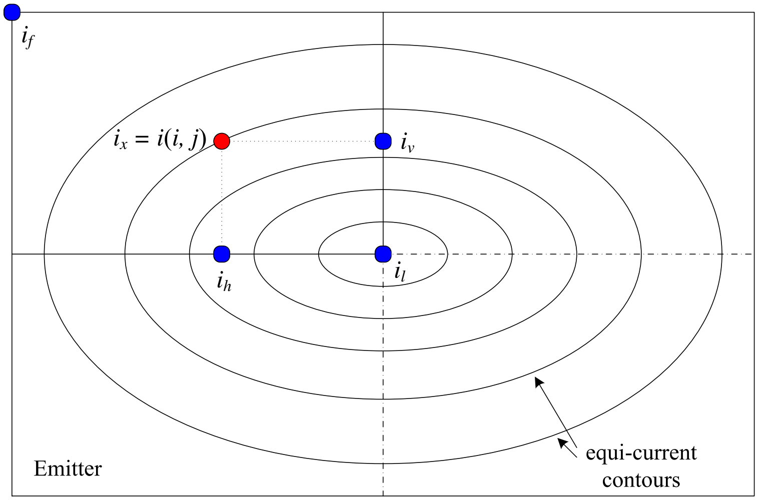

Figure 3.4 – Approximation of the inner transistor segment currents on the basis of four (center, corner, horizontal axis, vertical axis) full intrinsic transistor elements.

Mextram remained the same except for the part between nodes B_1 and B_2. This part (the resistor and the capacitor) models the internal distributed base resistance under the base-emitter capacitance. Since the model proposed in this chapter already has a built-in distributed transistor network, incorporation of these two elements would be redundant and incorrect, therefore they are excluded.

The decision on matrix dimensions, m and n, is on the user. Greater dimensions implicate higher precision, but also as a consequence the problem is more complex and it uses more resources, usually memory and processing time. In the implementation, assumption of four-fold rotational symmetry has been taken, yet this does not lessen the generality of the method which can even be applied for the multi-finger emitter geometries. Taking into account only one quadrant, the full intrinsic transistor elements are placed on the symmetry lines plus one in the corner while the interpolation transistor segments occupy all other positions in the matrix. The interpolation transistor segments are implemented in accordance with (3.1). Schematic of the interpolation transistor, composed of two controlled current sources, is given in Figure 3.6.

3.2.3 Extraction of additional model parameter

Beside standard bipolar transistor compact model parameters, there is only one additional parameter that has to be known in order to fully define the presented model. The extra parameter is the intrinsic base interconnection resistance, that connects the base nodes of the intrinsic transistor elements and interpolation segments. In practice, to optimize the precision, the model dimensions m and n should be proportional to the emitter length L_E and width W_E. When this is the case it yields the base interconnection resistance R_{BIC} extraction formula, which is indeed very similar to

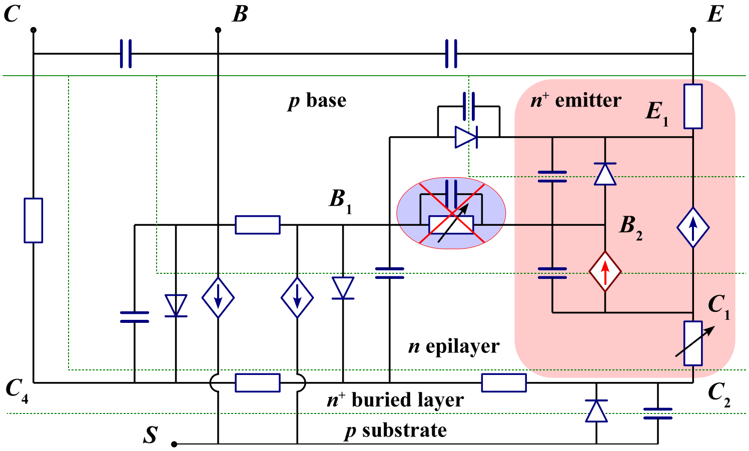

Figure 3.5 – The original Mextram model: Intrinsic transistor (highlighted) and obsolete part (struck-through) when the model with distributed base resistance is used.

the original estimation expression found in bipolar transistor compact models

$$R_{\mathrm{BIC}} = \rho_s \frac{m}{n} \frac{W_E}{L_E} \quad , \tag{3.2}$$

where ρ_s is the pinched sheet resistance of the base. Since the R_{BIC} models the resistance of equidistant and relatively uniform layers of the intrinsic base, all resistances in the resistive network of presented cases have the same value.

However, if there is a relatively large span in sheet resistance across the base region of a transistor, the assumption of equal resistances does not hold anymore and the resistances within the network should be updated accordingly. If the sheet resistance of the base is known at every point determining the network resistances is trivial.

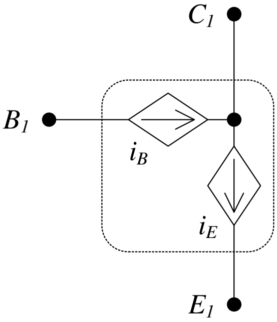

Figure 3.6 – The interpolation Segment: Inner transistor element replacement built of two controlled current sources that represent the base and the emitter current.

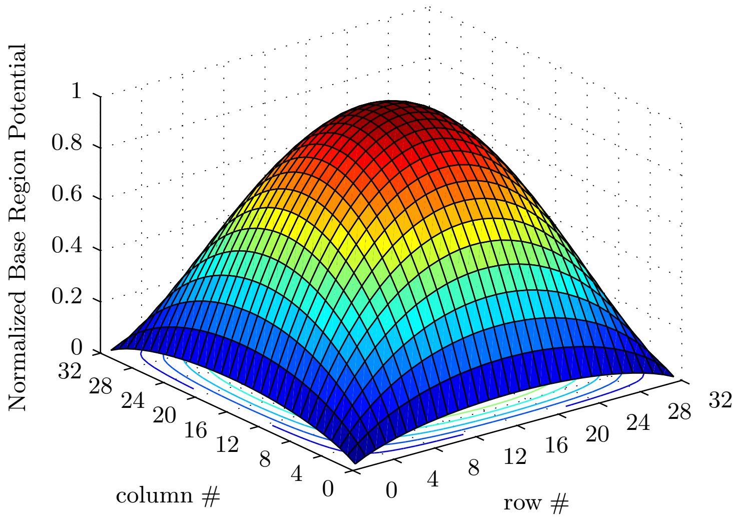

Figure 3.7 – Potential distribution the intrinsic base region (relative to emitter metal contact) obtained by segmented transistor model simulation in avalanche regime.

3.3 Simulation results of the sectionalized model

In practice, excellent match between measured and simulated data are obtained when the number of segments is larger than one thousand. Figure 3.7 shows the nonuniform normalized base electrostatic potential distribution that causes main current pinch-in. The simulation is performed on a square emitter geometry using a matrix with dimensions 32×32, model with 1024 segments. When symmetry is exploited, a model with only 256 segments remains. Then, the reduced model formulation employs only 32 full intrinsic transistor elements and 224 interpolation segments.

Maximum relative approximation error is in the order of magnitude of 0.1 percent. This relative error for approximated (in this case base and emitter) currents is drawn in Figure 3.8 as a function of applied DC voltage. The onset of impact ionization occurs around 12 V followed by the weak and strong avalanche breakdown.

Time consumption of the DC sweep simulations (for V_{CE} going from 0 to 20 V with a step of 0.1 V and fixed V_{BE}) with Synopsys HSPICE for two sectionalized transistors with different inner matrix dimensions is shown in Table 3.1.

The computation time is at that particular case halved in comparison to the original segmented model. The achievable gain in ideal situation should be a linear function of the matrix dimension. However, since the time consumption is mainly determined by calculation of the Jacobian matrix, and its size depends on the number of intrinsic transistor nodes that do not change, the computational gain in time is relatively modest. In practice, the computational complexity is expected to rise faster than linear but also slower than quadratic function of the input matrix dimensions.

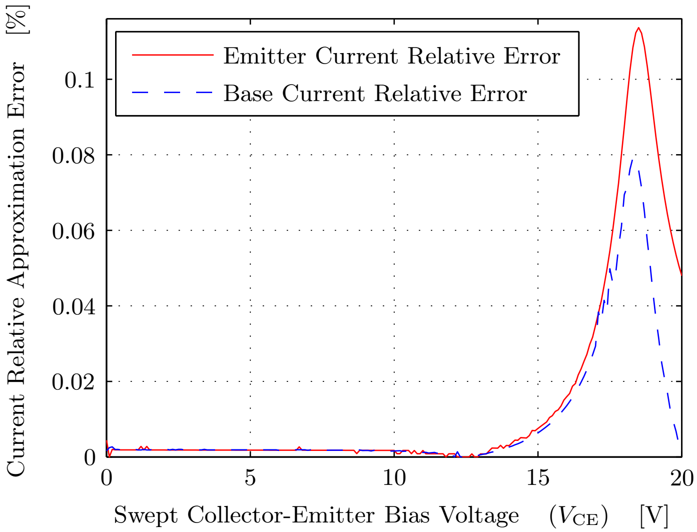

Figure 3.8 – Approximation error of the reduced model relative to the full model for the base terminal and the emitter terminal current in a DC sweep simulation.

3.4 Conclusion

A reduction technique for an accurate modeling of complex effects manifested in the avalanche regime of bipolar junction transistors is presented in this chapter. A phenomenon that is well-known but still not included in any of today's standard compact models is discussed first. Already existing quasidistributed three-dimensional transistor model is evaluated and its weakest point, the complexity, is emphasized. A simplification method that utilizes a normalized bilinear approximation is described. The rudiments of this method are explained in detail and the model is practically implemented in Verilog-A/AMS language. The model implementation is based upon Mextram. The additional parameter necessary for the full model definition is identified and its extraction technique is portrayed. The quantitative and comparison (with currently available model) results are discussed. The results are showing a significant gain in calculation time without notable loss in accuracy. Excellent agreement between the simulated and the experimental results is achieved.

Table 3.1 – Comparison of the computation (CPU) times for relative and reduced model.

	Reduced Model	Full Model
Matrix size of 20 × 20 elements	0.60 seconds	1.05 seconds
Matrix size of 32 × 32 elements	2.54 seconds	5.67 seconds

Chapter 4

Avalanche breakdown of bipolar transistors in AC regime

In the face of increasing demands for high frequency and high output power of modern bipolar transistor circuits, electronic circuit designers are exploring regimes of transistor operation that meet both requirements and enter RF regimes where impact ionization is significant. The present chapter addresses AC/RF avalanche characterization techniques. Repercussions of avalanche breakdown on some important transistor properties like unilateral power gain and Rollett's stability factor are introduced and demonstrated by measurements on modern industrial devices. On the basis of theoretical considerations and compact model simulations it is shown when avalanche can be expected to have significant impact on AC performance of transistors.

4.1 Introduction

The millimeter wave (mmW) bands offer exciting opportunities for various applications, such as short-range high data rate communications (e.g., the 60 GHz band), automatic cruise control and collision avoidance systems through automotive radars (e.g., the 77 GHz band) or passive imaging (e.g., the 94 GHz band) for security screening. Therefore, the research and development of silicon-based solutions for such mmW applications has gained significant momentum in recent years.

Perhaps the most challenging building block at mmW frequencies is a power amplifier (PA). Driven by the requirements in terms of costs, integration and performance, integrated circuit designers strive for implementation of such circuits using heterojunction bipolar transistors (HBT) available through the SiGe BiCMOS technology. Due to the mentioned tradeoff between transit time and breakdown voltage [24, 25], the speed improvement of modern SiGe(:C) processes has (partially) been achieved at the expense of reduced breakdown voltages. For bipolar junction transistors (BJTs) in general, device metrics such as the unity current gain bandwidth f_T, and unity power gain frequency $f_{\max}$, increase as the collector-base bias voltage V_{CB} (and collector-

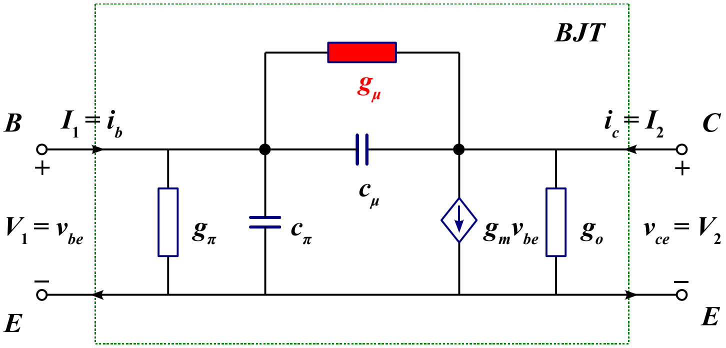

Figure 4.1 – The hybrid-pi model of a bipolar junction transistor used for small signal alternating current (AC) analysis complemented with the avalanche conductance g_μ, which is to be characterized and whose effects on the characteristics are to be analyzed.

emitter bias voltage V_{CE}) is increased. In the face of increasing demands set upon power amplifiers for high frequency and high output power, electronic circuit designers are exploring regimes of transistor operation that meet both requirements and enter RF regimes where impact ionization effects play significant role. Examples [73, 74] of power amplifiers implemented in SiGe:C BiCMOS technology and working in a $50 - 100\,\mathrm{GHz}$ range that exploit HBTs biased in the neighbourhood of the $\mathrm{BV_{CEO}}$ (the open base collector-emitter breakdown voltage) or even exceeding it by two to three [75] times, have been presented in the recent literature. Circuits that contain bipolar transistors that are operated above their $\mathrm{BV_{CEO}}$, but still always below $\mathrm{BV_{CBO}}$ (the base-collector p-n junction breakdown), are typically found in applications where high efficiency and large voltage swings are required, that is, in RF power amplifiers for mobile wireless applications. Specialized circuit topologies for biasing such circuits [76], tolerating output voltages above $\mathrm{BV_{CEO}}$ have also been reported.

For circuit designers that explore RF regimes in which avalanche breakdown can be expected to be significant, it is crucial to know the maximum usable transistor output voltage and its dependence on driving conditions, as well as the repercussions of working in the avalanche regime on other transistor properties like various figures of merit for power gain and stability. Integrated circuit designers rely on circuit simulator software to support their design process using computer simulations. Compact models, which accurately describe the behavior of transistors in a mathematical way are essential in these simulations thanks to their efficiency. Consequently, compact modeling of the AC/RF behavior of bipolar transistors in the impact ionization regime has become vital for the design of high-speed Si and SiGe bipolar circuits.

In general, studies of avalanche in bipolar transistors can be grouped in three classes, according to the regime in which a device is biased and the signal type by which it is driven. The first class would then be the static, direct current (DC) class, in which in fact no time-dependent signal is applied. This class is well-covered in

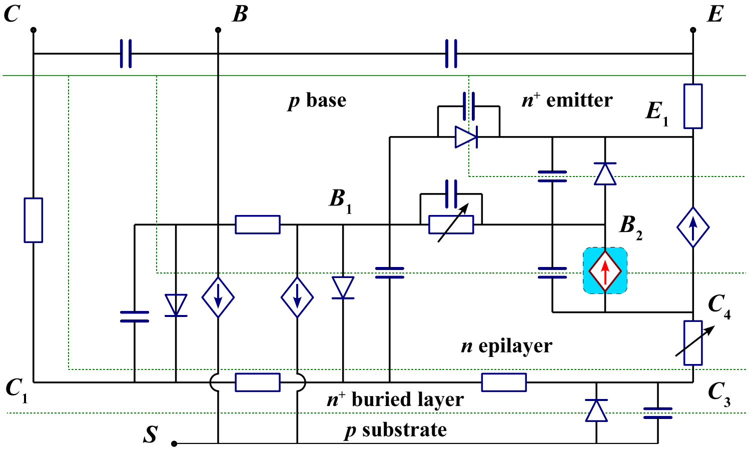

Figure 4.2 – Full equivalent circuit schematics of the bipolar transistor model Mextram 504.8. The avalanche current source, central to this chapter, is encased and highlighted.

literature [31], which presents models [47, 49, 50] that describe device behavior well, at least in the so-called weak avalanche regime. The second class could be referred to as small signal AC, where a device is *biased* within the avalanche regime and a small alternating signal is applied. In the third class a transistor is biased outside of the avalanche regime, but the applied AC signal is large enough to put the device into the avalanche regime during signal swings or vice versa.

The present chapter falls in the second class. It focuses on characterization of small signal AC bipolar transistor behavior in the impact ionization regime on radio frequencies (RF). The effects of avalanche on intrinsic transistor properties like unilateral power gain (G_U) and Rollett's stability factor (k) are addressed and analyzed on the basis of small signal equivalent circuit analysis and compact model simulations.

4.2 Small signal AC avalanche characterization

4.2.1 Theoretical considerations

In a standard hybrid-pi small signal equivalent circuit of a bipolar junction transistor (BJT) shown in Figure 4.1, the conductance g_μ in the base-collector junction due to avalanche is represented by a simple resistor $r_\mu = 1/g_\mu$ between the base and collector nodes. A straightforward hand calculation (which solves port current equations which are dependent on input voltages) shows that the avalanche conductance g_μ adds up to the real parts of all four two-port admittance parameters of this equivalent circuit

$$y_{11} = g_\pi + g_\mu + j\omega\left(c_\pi + c_\mu\right) \ , \quad y_{12} = -g_\mu - j\omega c_\mu \ , \tag{4.1}$$

$$y_{21} = g_m - g_\mu - j\omega c_\mu \ , \quad y_{22} = -g_o + g_\mu + j\omega c_\mu \ . \tag{4.2}$$

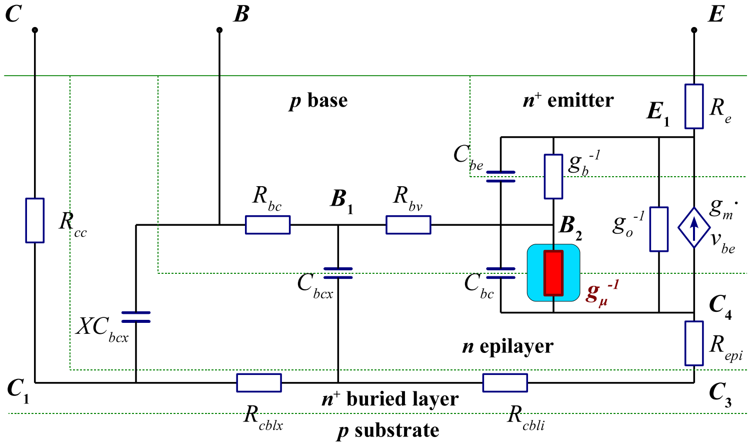

Figure 4.3 – Linearized equivalent circuit of the model of Figure 4.2, used in small signal analysis. Substrate coupling is neglected. The avalanche conductance is highlighted.

Imaginary unit is denoted by j. In this two-port network transistor representation, base-emitter nodes are considered to form port 1 and the collector-emitter nodes form port 2, hence together forming a common emitter configuration.

The real part of the y_{12} admittance parameter $\Re(y_{12})$ is equal to just $-g_\mu$. This suggests that in practice $-\Re(y_{12})$ might be an observable suitable to quantify and study avalanche in the small signals AC regime of transistor operation.

For accurate modeling of AC characteristics of modern industrial bipolar transistors in planar technologies, the hybrid-pi model of Figure 4.1 is too simple. The more extensive equivalent circuit, such as the one presented in Figure 4.3 with highlighted avalanche resistance, has been demonstrated [61] to be adequate to this aim. This circuit is a linear counterpart of the full equivalent circuit of Mextram [50] (shown in Figure 4.2 with highlighted avalanche current source), a physics-based standard compact model for vertical bipolar transistors. Compared to the hybrid-pi model in Figure 4.1, the circuit in Figure 4.3 takes parasitic resistances and capacitances in the emitter, base and collector of the transistor into account. To limit the complexity of further analysis of the circuit, collector-substrate coupling effects have been neglected.

Using a computer algebra system, it can be shown that the real part of the y'_{12}

Table 4.1 – Transistors on which Experimental Verification is Performed

	Transistor	Emitter [μm^2]	BV$_{\text{CEO}}$	Sweep	in Figures
I	QUBiC4X [77]	0.4×10.3	≈ 2.0 V	Frequency and Bias	4.4, 4.9, 4.12 4.5, 4.8, 4.11
II	Typical Industrial	0.4×0.8	≈ 5.5 V	Frequency	4.4′, 4.9′, 4.12′
III	QUBiC4X [77]	0.4×1.0	≈ 1.9 V	Bias	4.6, 4.7, 4.10

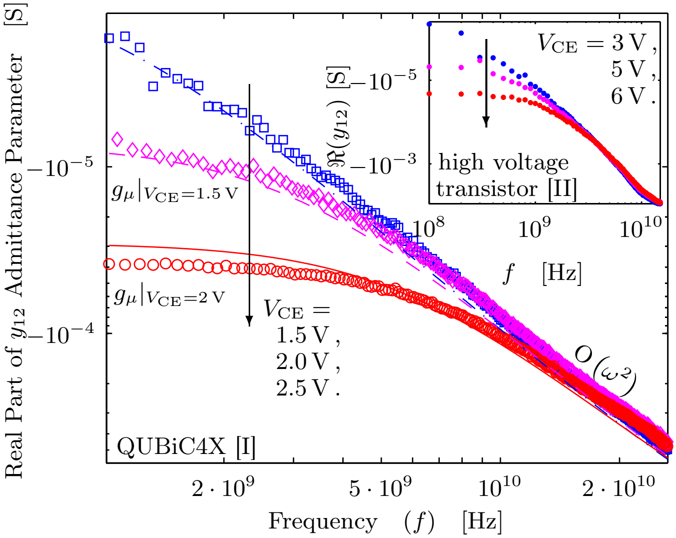

Figure 4.4 – Measured (markers) and simulated (lines) real part of y_{12} parameter as a function of frequency, for three values of the collector-emitter bias voltage (1.5 V squares and dash-dot line, 2.0 V diamonds and dashed line and 2.5 V circles and solid line) of device [I] that features $\mathrm{BV_{CEO}} \approx 2.0\,\mathrm{V}$. In the inset the corresponding measurements (dots) were performed on device [II] with $\mathrm{BV_{CEO}} \approx 5.5\,\mathrm{V}$ for collectore-emitter voltages of 3 V, 5 V and 6 V. On lower frequencies and higher voltages avalanche conductance is clearly dominant for both of the processes. The $\mathrm{O}\left(\omega^2\right)$ effects become dominant over g_μ as the frequency is increased, however, the turnover point occurs later on RF device.

admittance parameter of the linearized compact model two-port network circuit representation in common emitter configuration of Figure 4.3 is given by

$$-\Re\left(y'_{12}\right) = g_\mu + \mathrm{O}\left(\omega^2\right) \quad , \tag{4.3}$$

where $\omega = 2\pi f$ is angular frequency. The symbol $\mathrm{O}\left(\omega^2\right)$ denotes the additional terms of at least second order in ω. Because of space limitations, these terms cannot be represented here in full detail. Comparing the circuits from Figure 4.1 and Figure 4.3, and the corresponding resulting expressions (4.1) and (4.3), however, it can be easily shown that the $\mathrm{O}\left(\omega^2\right)$ terms represent the effects from parasitic resistances and capacitances (in combination with other model quantities like transconductance).

It should be emphasized that the effects of parasitics on $\Re(y'_{12})$ are of the second order as a function of frequency, as opposed to, for example, the first order contribution of base-collector capacitances to the $\Im(y_{12})$ in (4.1). On the basis of this observation, it can be expected that, if the measurements are to be taken at sufficiently low frequencies and sufficiently high output bias voltages, the real part of y_{12} is going to be dominated by the avalanche conductance g_μ. Nonetheless, at sufficiently high frequencies, the $\mathrm{O}\left(\omega^2\right)$ parasitic terms will dominate $-\Re(y'_{12})$.

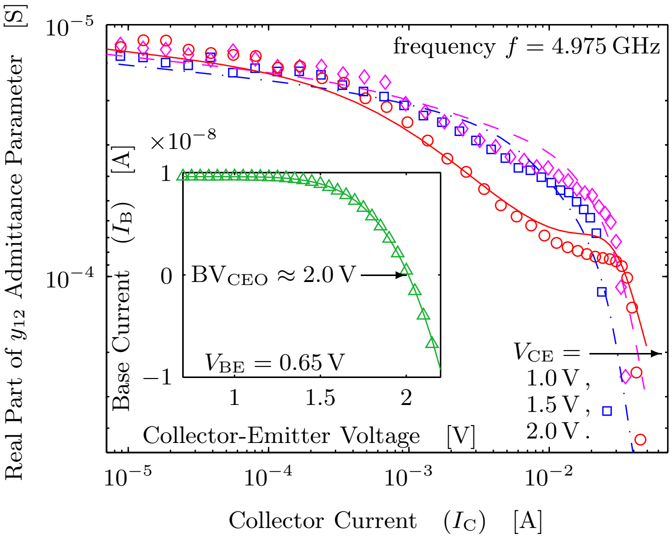

Figure 4.5 – Measured (markers) and simulated (lines) real part of y_{12} parameter as a function of collector current, for three values of the collector-emitter bias voltage (1.0 V squares and dash-dot line, 1.5 V diamonds and dashed line and 2.0 V circles and solid line). In the subplot forward Early measurement (triangles) and simulation (solid line) of the base current versus applied collector-emitter voltage are shown to indicate an avalanche breakdown point. Measurements correspond to a device [I]. In the medium current region, a complete change in trend of the $\Re(y'_{12})$ admittance parameter (the 2.0 V curve crosses 1.0 V and 1.5 V curves) caused by avalanche is observed.

4.2.2 Experimental verification

The qualitative behavior of the real part of y_{12} admittance parameter as a function of frequency and bias conditions, as expected from the theoretical considerations of the previous subsection, is indeed clearly observed in the measured data that was taken from three representative, modern industrial SiGe:C heterojunction bipolar transistors. The most relevant properties of these HBTs are summarized in Table 4.1. Two high speed (HS) QUBiC4X BNX-type BiCMOS HBTs and one typical industrial high voltage (HV) BiCMOS SOI HBT are used. All transistors are of NPN type.

The measured (symbols) frequency sweeps in the main plot of Figure 4.4, as well as of Figure 4.9 and Figure 4.12, are the ones of device [I] presented for several bias conditions: ($V_{BE} = 0.82$ V, $V_{CE} = 1.5$ V), ($V_{BE} = 0.81$ V, $V_{CE} = 2.0$ V) and ($V_{BE} = 0.81$ V, $V_{CE} = 2.5$V). In the insets of the same figures, frequency sweeps of another device [II] are presented for $V_{CE} = 3$, 5, and 6 V, all for $V_{BE} = 0.85$ V. These, and all other measurements in this chapter are taken at room temperature. Also all RF measurements are deembedded using the on-wafer open and short structures for high speed devices and only on-wafer open structure for high voltage device.

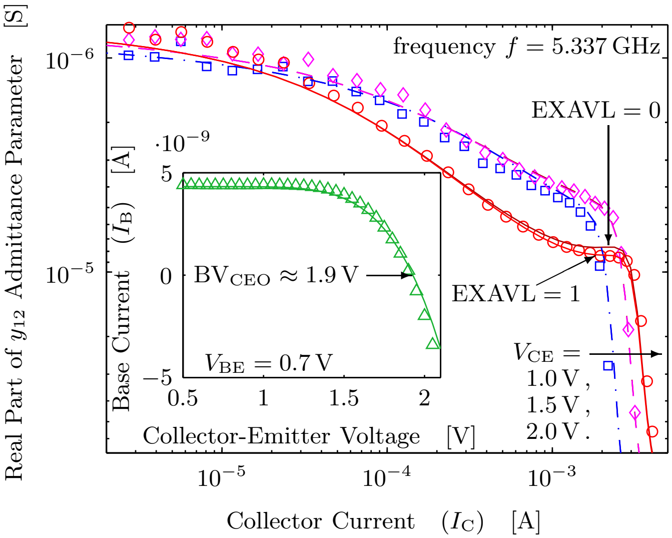

Figure 4.6 – Measured (markers) and simulated (lines) real part of y_{12} parameter as a function of collector current, for three values of the collector-emitter bias voltage (1.0 V squares and dash-dot line, 1.5 V diamonds and dashed line and 2.0 V circles and solid line). In the subplot forward Early measurement (triangles) and simulation (solid line) of the base current versus applied collector-emitter voltage are shown to indicate an avalanche breakdown point. Measurements correspond to a device [II]. In the medium current region, a complete change in trend of the $\Re(y'_{12})$ admittance parameter (the 2.0 V curve crosses 1.0 V and 1.5 V curves) caused by avalanche is observed.

For V_{CE} well below $\mathrm{BV}_{\mathrm{CEO}}$, it can be observed the $\mathrm{O}\left(\omega^2\right)$ dependence of $\Re(y'_{12})$ for all frequencies within the measurement range. This corresponds to a frequency dependence of $\Re(y'_{12})$ induced by parasitic resistances and capacitances, conforming (4.3). For both devices it is observed that in the appropriate low frequency limit, for V_{CE} approaching or above $\mathrm{BV}_{\mathrm{CEO}}$, $\Re(y'_{12})$ tends to a frequency independent value. Apparently it is dominated by g_μ in this regime. In the high frequency limit, for the higher V_{CE} values shown, parasitic effects are observed to dominate avalanche conductance.

Figure 4.5 presents the bias sweep measurement (symbols) at fixed frequency of $f = 4.975$ GHz for the same QUBiC4X BNX-type transistor. Measurements shown in Figure 4.8 and Figure 4.11 are obtained on the very same transistor as well.

For comparison purposes, one more QUBiC4X device [III] for bias sweeps is taken with only the emitter size different. Figure 4.6, shows the measured (symbols) data that are obtained from this device, as well do the Figure 4.7 and Figure 4.10.

It can be observed that in the medium current region, the effects of bias cause a global increase of the value of $|\Re(y'_{12})|$. Figures 4.5 and 4.6 also show a general increase of $|\Re(y'_{12})|$ with increasing I_{C}, due to the bias-dependence of elements of the equivalent circuit. The drastic increase in $|\Re(y'_{12})|$ for I_{C} above roughly 2 mA in

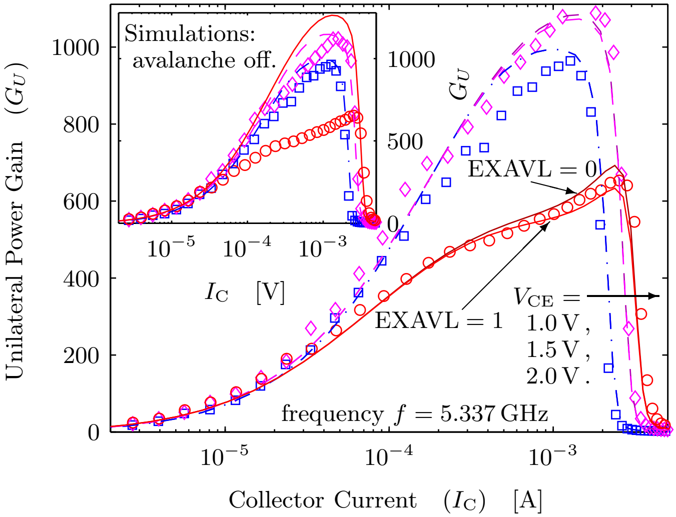

Figure 4.7 – Measured (markers) of [III] and simulated (lines) unilateral power gain as a function of collector current, for three values (1.0 V, 1.5 V and 2.0 V) of the collector-emitter bias voltage V_{CE}. Symbols and curves correspond to those in Figure 4.5 and Figure 4.6. In the inset the same plot is repeated with the avalanche model of Mextram turned off in device simulations, thus giving a clear overestimate of the power gain for $V_{CE} = 2.0$ V, where impact ionization effects cannot be neglected.

smaller and above 20 mA in larger device is due to the so-called base pushout or Kirk effect. It is well-known that base pushout is postponed when V_{CE} is increased and this trend is clearly observed in the figures. The effect of avalanche on $|\Re(y_{12})|$ is clearly recognizable in the collector current I_C regime between roughly 30 uA to 1.5 mA, as it causes a dramatic increase in magnitude when V_{CE} is increased from 1.5 V to 2.0 V. All the measurements in this chapter were taken at room temperature.

4.3 Small signal AC avalanche modeling

The curves in Figure 4.5 and Figure 4.6 demonstrate the capability of Mextram 504.8 to simulate effects of avalanche on AC characteristics. The inset of the figure shows measured (symbols) and simulated (curve) data under DC bias conditions [54], more specifically, the DC base terminal current I_B for fixed V_{BE}, as a function of V_{CE}. The observed decrease of base current due to avalanche is neatly reproduced by the model simulations. The main plot of Figure 4.6 demonstrates that the combination of Mextram's (version 504.8) quasistatic avalanche model [56] and its extensive modeling of parasitic effects [61] provides the capability to accurately simulate the effects of impact ionization on the shown RF characteristics as well. Here also extended

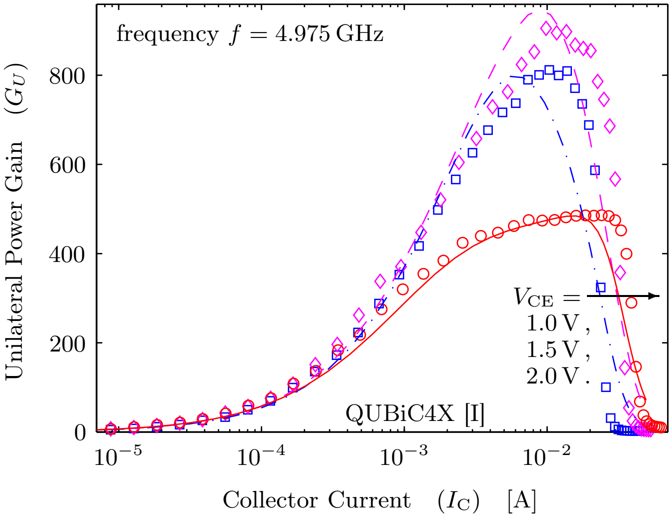

Figure 4.8 – Measured (markers) of [I] and simulated (lines) unilateral power gain as a function of collector current, for three values (1.0 V, 1.5 V and 2.0 V) of the collector-emitter bias voltage V_{CE}. Symbols and curves correspond to those in Figure 4.6.

avalanche model is turned on by switching EXAVL flag from 0 to 1. It enables taking into account the effective epilayer width decrease due to base widening. The effect of the extended impact ionization modeling is observable in Figure 4.6 and Figure 4.7, which show that in the critical region of transition between middle to high current regime, setting the flag EXAVL = 1 gives better fit.

In Figure 4.4, it is shown that the dependence on frequency of $|\Re(y'_{12})|$ can be simulated (curves) well by Mextram. The underestimation of $|\Re(y'_{12})|$ in the low frequency limit for the $V_{\mathrm{CE}} = 2.5$ V curve is due to the intentional underestimation of avalanche in the strong avalanche regime. Indeed, for the reasons of robustness of convergence in the context of general circuit simulations, the avalanche model of Mextram is restricted to weak avalanche.

4.4 Small signal AC avalanche repercussions

Expressions (4.1) and (4.2) show that in a two-port description of the small signal behavior of a bipolar transistor, avalanche is manifested in the real part of all admittance parameters. In turn, the real parts of admittances y are crucial ingredients of some fundamental properties or invariants, of two-ports. In this section the impact of avalanche on some of these is explored, namely on the unilateral power gain G_U, Rollett's stability factor k and the maximum available power gain G_{MA}.

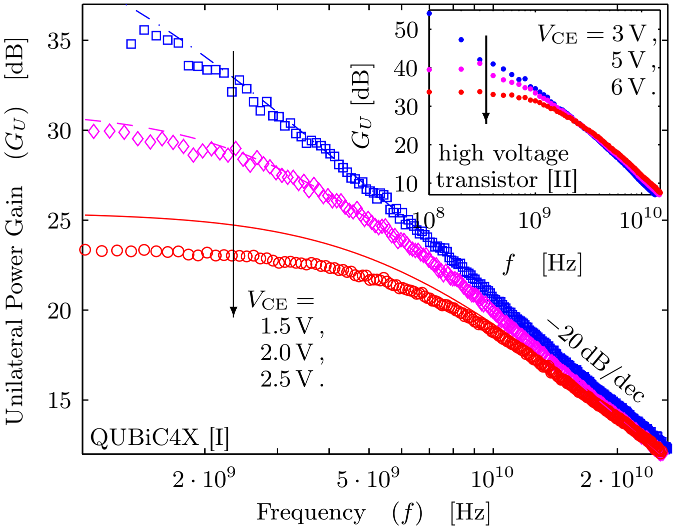

Figure 4.9 – Measured (markers) and simulated (lines) unilateral power gain as a function of frequency of the transistor [I], for three values (1.5 V, 2.0 V and 2.5 V) of the collector-emitter bias voltage. In the inset the corresponding measurements (dots) were performed on the transistor [II] for three values (3 V, 5 V and 6 V) of V_{CE} bias. Clear drop of the gain caused by a negative feedback loop effect caused by the avalanche conductance on higher output voltages V_{CE} and lower frequencies is observed.

4.4.1 Unilateral power gain

As an invariant and hence an intrinsic device property, unilateral power gain [78] is a central concept in two-port active device characterization [79], as well as in (RF) analog circuit design. It is expressed in terms of two-port network parameters by the following equation

$$G_U = \frac{|\gamma_{21} - \gamma_{12}|^2}{4\left(\Re(\gamma_{11})\Re(\gamma_{22}) - \Re(\gamma_{12})\Re(\gamma_{21})\right)} \quad , \tag{4.4}$$

where the immittances γ can be substituted by impedance- (z), admittance- (y), hybrid- (h) or inverse hybrid (g) parameters of the two-port network transistor representation. As shown by expression (4.4), the unilateral power gain explicitly depends on the real parts of all four two-port transistor representation parameters.

Equation (4.2) shows that, at lowest order in ω, $\Re(y_{21})$ is the intrinsic transconductance of the transistor, g_m. Since in the forward active mode of transistor operation, g_m is a dominant quantity in bipolar transistors, and in view of the findings in the previous sections about the sensitivity of $\Re(y_{12})$ to avalanche, (4.4) suggests that the unilateral power gain may be sensitive to avalanche too. Such sensitivity can indeed be observed on modern industrial bipolar transistors, as is demonstrated in Figure 4.7,

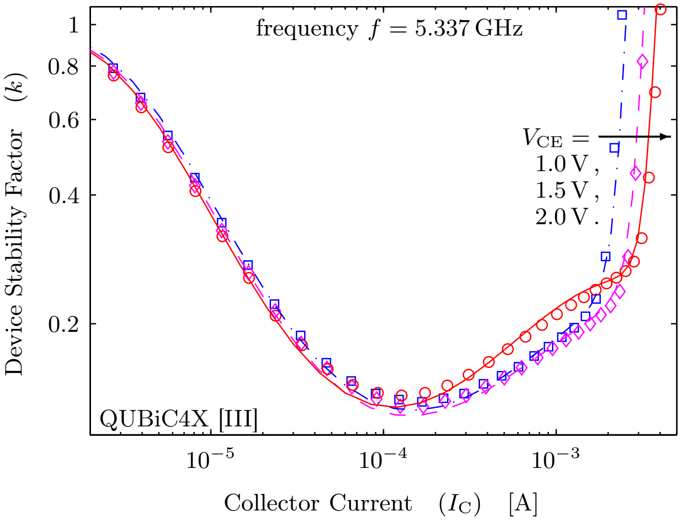

Figure 4.10 – Measured (markers) of [III] and simulated (lines) stability factor k as a function of collector current, for three values (1.0 V, 1.5 V and 2.0 V) of the collector-emitter bias voltage V_{CE}. Symbols and curves correspond to those in Figure 4.6 and Figure 4.7. Also a complete change in trend of the observed variable may be noticed in middle current range, as is the case for $\Re(y_{12})$ and G_U, of mentioned figures, respectively.

Figure 4.8 and Figure 4.9. These figures show the measured (symbols) and the simulated (curves) unilateral power gain as a function of bias at fixed frequency and as a function of frequency at several biases, respectively.

In Figure 4.7 and Figure 4.8 it can be observed that for $V_{\mathrm{CE}} = 1.0\,\mathrm{V}$ or $1.5\,\mathrm{V}$, at fixed collector current I_{C}, the unilateral power gain increases with increasing V_{CE}. When V_{CE} is further increased to 2 V however, the gain drops dramatically. The physical mechanism behind this sensitivity is demonstrated in Figure 4.7 by means of compact model simulations using Mextram. The curves in the main plot of this figure show the result of Mextram simulations. The curves in the inset of the figure also show the results of Mextram simulation, but with a model parameter set that effectively suppresses avalanche effects altogether. These simulations clearly identify avalanche as the cause of the dramatic drop of G_U when V_{CE} is increased to 2.0 V. As a general observation, it should be noted that the underestimation of avalanche leads to the overestimation of the transistor's (unilateral) power gain at some frequencies.

Also in Figure 4.7 and Figure 4.8 a sudden drop of unilateral power gain may be observed at high current values. As this sharp change occurs at the very same current level as for $\Re(y'_{12})$ it might be expected that the same phenomenon is responsible for this and that it is also the base pushout. However, it may be observed in Figure 4.5 and in Figure 4.6 that there is a change in observed variable for fixed output voltage. For this change $\mathrm{O}\left(\omega^2\right)$ effect is responsible, as it includes bias dependent terms.

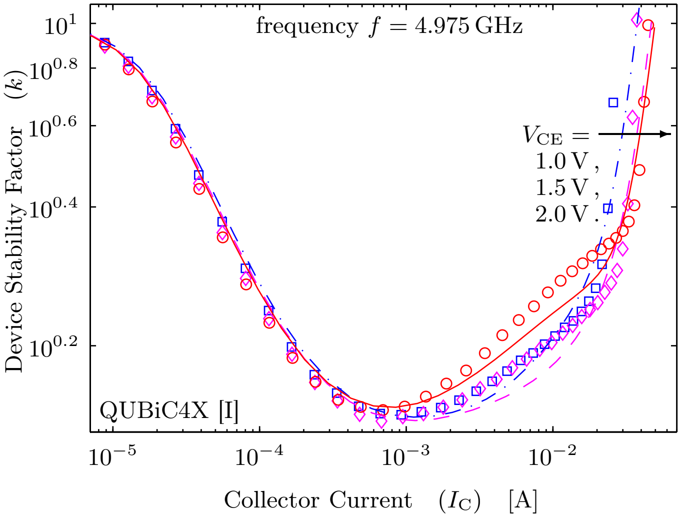

Figure 4.11 – Measured (markers) of [I] and simulated (lines) stability factor k as a function of collector current, for three values (1.0 V, 1.5 V and 2.0 V) of the collector-emitter bias voltage V_{CE}. Symbols and curves correspond to those in Figure 4.5 and Figure 4.8. Also a complete change in trend of the observed variable may be noticed in middle current range, as is the case for $\Re(y_{12})$ and G_U, of mentioned figures, respectively.

In Subsection 4.2.2 we learned that at sufficiently high frequencies, $\Re(y'_{12})$ may be dominated by the parasitic effects that in such cases mask the influence of avalanche. These masking frequencies are higher as the avalanche conductance is larger.

The conductance $-\Re(y_{12})$ represents a feedback inside an amplifying device. Due to the phase rotation of π between the intrinsic collector and base nodes this feedback is negative, therefore the minus sign. As a result, power gain will drop with frequency, as follows from the feedback theory [10], when $-\Re(y'_{12})$ increases with frequency due to the parasitic base and collector resistance effects and their distribution over the base-collector capacitance, represented by the $\mathrm{O}\left(\omega^2\right)$ term in (4.3). For bias conditions that are outside the avalanche regime ($V_{\text{CE}} = 1.5\,\text{V} \ll \text{BV}_{\text{CEO}}$ for HS and $V_{\text{CE}} = 3.0\,\text{V} \ll \text{BV}_{\text{CEO}}$ for HV), this frequency dependence is indeed observed in Figure 4.9, which presents the measured (symbols) and Mextram simulated (curves) values of G_U as a function of frequency; results are shown for the same two representative industrial SiGe BiCMOS HBTs as in Figure 4.4. The same figure shows that, similar to $-\Re(y'_{12})$ (Figure 4.4), in the low-frequency limit, G_U is sensitive to avalanche effects, while in the high frequency limit, this effect is masked by the (parasitic) resistance-capacitance effects. In the low frequency limit and for $V_{\text{CE}} = 2.5\,\text{V}$, the impact of avalanche on G_U is underestimated by Mextram 504.8 due to the intentional restriction of the avalanche model to weak avalanche (see also the remarks at the end of Section 4.3).

A question of general interest is the following: up to which operating frequency are

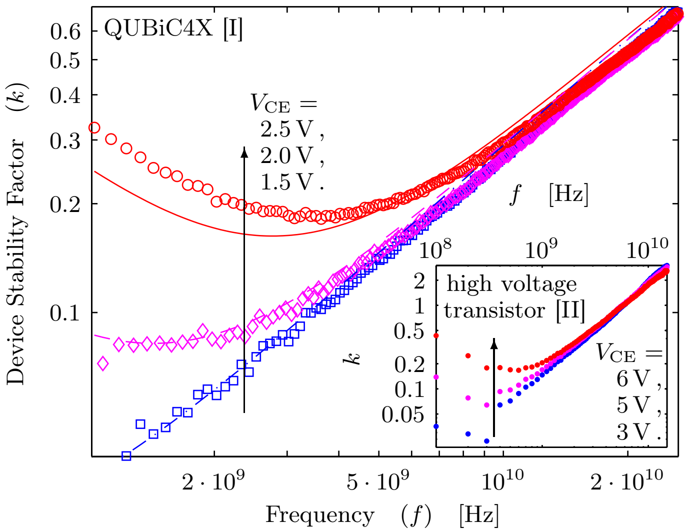

Figure 4.12 – Measured (markers) and simulated (lines) frequency response of the stability factor k of the high speed RF device [I], for three values (1.5 V, 2.0 V and 2.5 V) of the collector-emitter bias voltage V_{CE}. In the inset the corresponding measurements were performed on a high voltage device [II] for three values (3 V, 5 V and 6 V) of V_{CE}. Drastic increase of stability caused by the avalanche conductance feedback effect can be observed in the lower frequency region when it enters into the impact ionization regime. Symbols and curves correspond to those in Figure 4.4 and Figure 4.9.

the avalanche effects actually dominant over the (distributed) parasitic resistance and capacitance effects? As explained in Subsection 4.2.1, the second order effects in (4.3) strongly depend on the distributed parasitic base-collector capacitances. As a result the frequency above which the unilateral power gain will start to fall off with increasing frequency will depend on these capacitances. The lower the capacitances will be, the higher will be the mentioned frequency. Therefore, as a result of the industrial trend of decreasing the parasitic base-collector capacitance in order to increase the maximum oscillation frequency f_{max}, avalanche effects can be expected to become more and more pronounced in RF characteristics and figures of merit of future high speed technology generations (as seen by quantitative comparison of devices [I] and [II] in this aspect).

Measuring at mmW frequencies can be both challenging and expensive, therefore it would be useful to be able to make a relatively accurate estimate of the frequency up to which the impact ionization effects are dominant over the higher order parasitic resistance and capacitance effects (and actually prevail) in the G_U characteristics. As such estimate one might choose the frequency point where −20 dB/dec unilateral gain extrapolation of the device, as biased with output voltage which does not correspond to excessive impact ionization (low V_{CE}) with input bias voltage kept at the same value, would meet the measured avalanche affected unilateral gain at low frequencies.

4.4.2 Stability factor

Other important invariants, useful in RF circuit design [80], are the stability factor,

$$k = \frac{2\Re(\gamma_{11})\Re(\gamma_{22}) - \Re(\gamma_{12}\gamma_{21})}{|\gamma_{12}\gamma_{21}|} \quad , \tag{4.5}$$

and the maximum available power gain G_{MA},

$$G_{MA} = \left|\frac{\gamma_{21}}{\gamma_{12}}\right| (k - \sqrt{k^2 - 1}) \quad . \tag{4.6}$$

In the last expressions the immittances γ can be substituted by impedance- (z), admittance- (y), hybrid- (h) or inverse hybrid (g) transistor two-port network representation parameters, meaning k and G_{MA} are invariants.

The measurements (symbols) and the model simulations (curves) of the stability factor as a function of bias at fixed frequency and as a function of frequency are shown in Figure 4.10, Figure 4.11 and Figure 4.12, respectively. The influence of avalanche on the maximum available gain is similar to its influence on the unilateral power gain (of course, when defined for $k \geq 1$). In fact, when avalanche influence is underrated, the maximum available power gain is overvalued. Repercussion of impact ionization on stability can also be explained in terms of the avalanche-induced negative feedback. Namely negative feedback tends to move transfer function poles further away from the right complex half-plane [10], that is, the device becomes more stable.

4.5 Conclusion

To meet the increasing demands for high operating frequency and high output power in modern bipolar transistor applications, circuit designers explore regimes of transistor operation close to or within the avalanche breakdown region.

In order to qualify and quantitatively model the effects occurring in the impact ionization regime, in the present chapter this regime of operation is addressed, in which small signal bipolar transistor behavior is analyzed. The collapse of the unilateral power gain due to the impact ionization effects, as quantified by the avalanche-induced conductance g_μ, is demonstrated, physical origin of it is identified and the repercussions of avalanche on the maximum available power gain, as well as on Rollett's invariant, the stability factor k, is addressed. The frequency dependence of these quantities is described and commented in detail. The concepts and analyses are illustrated by the RF measurements on modern industrial heterojunction devices and by the corresponding computer simulations employing the standard compact model for bipolar transistors. Though all examples in the chapter were performed measuring and simulating NPN types, all physical concepts are qualitatively and quantitatively also applicable to PNP type bipolar junction transistors. It is found out that the effects of avalanche on AC characteristics and figures of merit may be masked by higher order effects of parasitic resistances and capacitances. However, according to the conducted analysis, trends in industry imply that avalanche effects tend to be dominant over parasitic effects in the most modern and coming technology generations.

Chapter 5

Compact modeling of tunneling breakdown

PHYSICS-BASED compact model of band-to-band tunneling current in p-n junctions is presented in this chapter. The model features a smooth transition to zero forward bias tunneling current, full physical temperature scaling and an innovative parametrization. An accurate experimental verification of the physical temperature scaling rules for band-to-band tunneling is presented, on carefully selected state of the art industrial transistors. By simulations of statistics it is explicitly demonstrated, that the novel choice of model parameters yields improved parameter determinability. Furthermore, it is explicitly shown on measured data that this improved parameter determinability is essential for good geometrical scalability of the parameter values.

5.1 Introduction

A general trend in integrated circuit technologies, both with respect to CMOS and bipolar junction (BJT) transistors, is an increase of doping concentrations.

The main leakage mechanisms in CMOS, gate, junction and sub-threshold leakage, increase as field effect transistors are scaled down towards 10 nm [66]. In state of the art CMOS technologies, for example, the leakage power has become a significant portion of the total power dissipation. Therefore, it is imperative for circuit designers and system architects to have available accurate predictions of the system's leakage mechanisms [65]. The junction leakage in current CMOS generations is mainly due to the pocket implants, also called (super-)halos. These are used to combat short-channel effects [67] and off-state leakage. It has been suggested that *junction leakage* will present the fundamental limit for scaling of the traditional MOS transistor structure [36]. This leakage of reversely biased p-n junctions, Figure 5.1, so natural for drain-to-body junctions [68], is induced mainly by Shockley-Read-Hall (SRH) generation / recombination and trap-assisted tunneling (TAT) at relatively low voltages and higher temperatures, by band-to-band tunneling (BtBT) [81] in the middle voltage

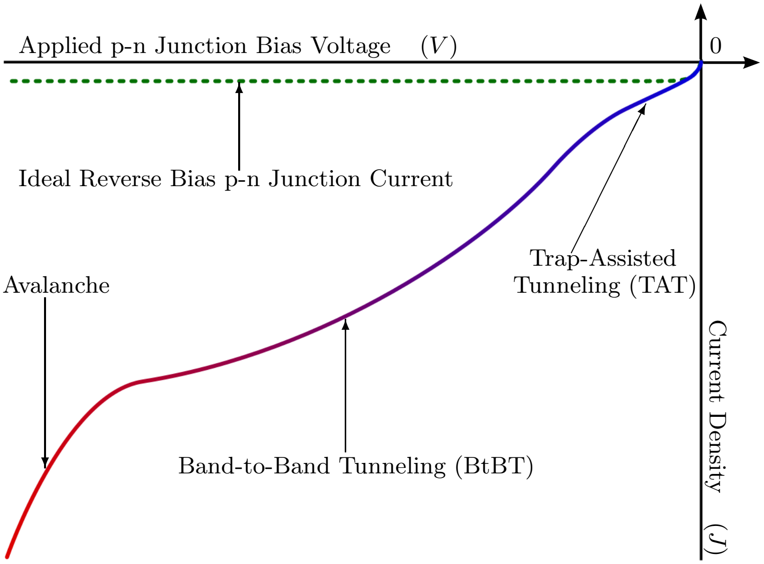
Applied p-n Junction Bias Voltage (V)
0
Ideal Reverse Bias p-n Junction Current
Avalanche
Trap-Assisted Tunneling (TAT)
Band-to-Band Tunneling (BtBT)
Current Density (J)

Figure 5.1 – Components of p-n junction reverse bias leakage current density. The ideal reverse bias current is also sketched. The current is dominated by trap-assisted tunneling (TAT) at low biases, by band-to-band tunneling (BtBT), at middle range biases and avalanche multiplication current at relatively high biases.

range, and finally by impact-ionization avalanche current at high voltages. Due to its presence in a wide voltage range of interest, accurate modeling of BtBT currents is an important issue in the overall process of dealing with transistor leakages [82].

Application of SiGe heterojunction bipolar transistor (HBT) technology in low noise and power amplifier stages for Wi-Fi WLAN, WiMAX and UMTS W-CDMA wireless communication, has brought with it a trend to ever higher doping concentrations, to increase the speed and (in a tradeoff) increase dynamic range, reduce power consumption and lower noise contribution of the transistors. High doping concentrations, as applied in the base-emitter junction of modern heterojunction bipolar devices, induce high electric fields and therefore may imply significant band-to-band tunneling (BtBT) currents in the reverse bias regime. For this reason, modern compact transistor models need to be capable to represent these currents.

Because, to meet demands for low power consumption for battery supplied applications, bipolar transistors are more and more applied at very low bias conditions, compact models increasingly need to be accurate in the low forward bias regime. Due to increasing bipolar transistor applications in extreme environments [83], most notably at cryogenic temperature conditions [84], models are increasingly expected to be applicable at very low temperatures as well.

Besides being faithful and accurate over a wide temperature range, semiconductor device compact models for semiconductor circuit simulations need to be efficient in terms of computational load. Furthermore, the parameters of the compact model should be reproducibly extractable from measured data, so as they will be scalable in terms of temperature and geometry.

In this chapter a physics-based compact model for p-n junction BtBT currents is presented that has been developed to address all these demands. The model features an identically vanishing forward bias Zener (band-to-band) tunneling current. This simplifies evaluation of the model, and hence enables reduction of the computational load, under forward bias conditions. Special attention has been paid to ensure a smooth transition from reverse to forward bias. The physical basis of this is discussed in Section 5.2.2 while the technical implementation is presented in the Appendix B.

Experimentally obtained data of band-to-band tunneling currents are not always reproducible and hence are not always reliable for accurate model verification. It may be for this reason that published models in the semiconductor literature [85, 55] have adopted various, mutually different, approximations of the full physical temperature scaling rules for BtBT currents. In Section 5.4, the merits of the fully physics-based temperature scaling model for band-to-band tunneling current shall be demonstrated on reproducible data obtained from a state of the art industrial bipolar device.

A third aspect in which the proposed model differs from earlier published models is in the definition of the model parameters. In Section 5.2.4, by numerical statistical simulation it is demonstrated that the sensitivity of the extracted parameter values to stochastic errors in measured data is strongly dependent on the parametrization of the BtBT model. It is shown that the novel parametrization reduces this sensitivity, as compared to a more traditional parametrization. Moreover, I demonstrated on experimental data taken from an in-house DIMES04 bipolar Si process that this parametrization greatly improves geometrical scalability of model parameter values.

5.2 Compact model foundations

5.2.1 General theoretical background

The presented compact model of band-to-band tunneling (BtBT), schematically depicted in Figure 5.2, is based on analytical formulations as documented in the semiconductor device physics literature. These are of the form

$$J_{\mathrm{BBT}} = \sigma \, \frac{\sqrt{m^*}}{{E_g}^{\nu}} \, E^{\xi} \, D \, \exp\left(-\theta \, \frac{\sqrt{2m^* {E_g}^3}}{q \, \hbar \, E} \right) \quad . \tag{5.1}$$

In this expression, J_{BBT} denotes the band-to-band tunneling current density, $\hbar$ is the reduced Plank constant, q the elementary charge and m^* the electron's effective mass. In applications to p-n junctions, E is to be taken as the maximum electric field [86]. The powers ν and ξ that partly describe the dependence of J_{BBT} on the band gap E_g and on the electric field E, depend on the physical origin [87] of the BtBT current. For direct BtBT $\nu = 1/2$ and $\xi = 1$, whereas for phonon-assisted tunneling $\nu = 1/4$ and $\xi = 3/2$. The numerical prefactor σ also depends on the precise physical origin of the tunneling current [88, 89]. The dimensionless numerical

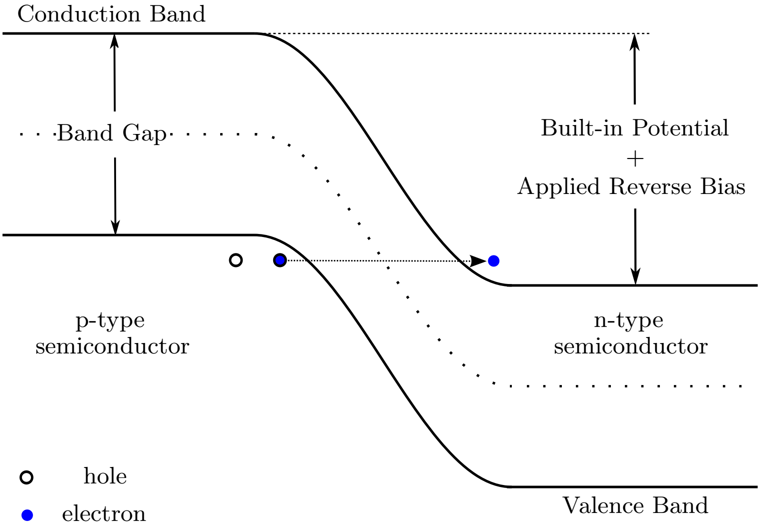

Figure 5.2 – Schematic representation of the band-to-band tunneling process in p-n junctions. A valence band electron from the p-type semiconductor tunnels through the forbidden band gap to the conduction band of the n-type material, leaving a hole behind. Such generated valence band holes and conduction band electrons are charge carriers of the junction's band-to-band tunneling (BtBT) current.

constant θ takes the value $\theta = 4/3$ if the tunneling probability is calculated for a triangular potential barrier [90], while it takes the value $\theta = \pi/4$ if the calculation is performed on basis of the representation of the forbidden gap by a parabolic potential barrier [88]. On basis of experiments with a full implementation of equation (5.1) and in accordance with reported findings in the literature [86], I have come to the conclusion that, in the context of compact modeling of p-n junctions, the coefficients ν and ξ are not well determined in terms of measured device characteristics, that is, the values of these coefficients cannot be extracted from observed characteristics and may, without significant loss of modeling capability, be taken as the computationally attractive $\nu = 1/2$ and $\xi = 1$, corresponding to direct tunneling. The coefficients σ and θ will be absorbed in compact model parameters.

5.2.2 State occupation factor and vanishing tunneling current

The band-to-band tunneling current in highly doped p-n junctions may be significant when the junction is forced in reverse bias. As long as the semiconductor material involved is nondegenerate (or even in which the built-in potential is lower than the band gap) however, tunneling will not contribute to the forward current, because final states of the tunneling process would correspond to the states in the forbidden energy gap [89, pg. 373]. Indeed, tunneling is only possible in the space charge region of p-n junctions when there are occupied initial states on one side and empty final states on the other side of the junction. This effect of state occupation is captured in expression (5.1) by the factor D, which in reverse bias can be approximated [91] by

$$D = -qV - \bar{E}(V)\left[1 - \exp\left(\frac{qV}{\bar{E}(V)}\right)\right] \quad , \tag{5.2}$$

where, V is the applied p-n junction bias voltage (negative for reverse bias) and

$$\bar{E}(V) \equiv \frac{\sqrt{2}}{\pi}\frac{q\,\hbar\,E(V)}{\sqrt{m^*\,E_g}} \quad , \tag{5.3}$$

is, as defined [91, eq. (11)][88, eq. (12-37)], a measure of the significant range of perpendicular momentum.

Expression (5.2) is found [91] in the case of direct band-to-band tunneling in (i) the limit of vanishing absolute temperature and (ii) the limit in which the electron and hole quasi-Fermi levels coincide with the conduction and valence band edges, respectively. In this limit, the tunneling current vanishes as a function of bias voltage V exactly at zero bias, $V = 0$. Tunneling effects under forward bias conditions also vanish in this limit. On the results for the tunneling current in the regime of relatively high reverse bias, the adoption of expression (5.2) has negligible influence. Presented compact model of BtBT aims to implement an accurate, physics-based description of the BtBT current in the regime where this current is significant, that is in the regime of relatively high reverse bias. In addition, I choose not to include forward tunneling effects. Last, the aim is to implement a smooth transition from the regime of tunneling current, into the regime of vanishing tunneling current. To meet these three requests, the form (5.2) proves to present a suitable starting point. How this expression serves the third aim is discussed in detail in the remainder of this section.

In practice, the factor D is commonly [85, 55] approximated by its leading term $-qV$. With respect to the value of the tunneling current, this is well-justified whenever the tunneling current adds a significant contribution to the reverse bias current, because for sufficiently large voltages V the leading term qV is dominant over the second term in expression (5.2). However, this approximation significantly changes the derivatives of D and hence of J_{BBT}, with respect to applied voltage V, at zero bias. Indeed, the first derivative of D, as defined by (5.2), with respect to V, vanishes at $V = 0$, whereas the derivative dJ_{BBT}/dV of its approximation by $-qV$ does not. This is demonstrated in Figure 5.3, in which a model that adopts leading term approximation of D is represented by the dashed curve. As the figure shows, the model predicts a finite overshoot-like shape of the tunneling current under forward bias conditions. When obtained in this way though, this prediction is not physical. To implement a physical description of forward tunneling, a detailed physical representation of the D factor [91] should be included in the formulation. If the vanishing of the tunneling current under forward bias would be implemented by just setting the forward current to zero, while maintaining the approximation $D \approx -qV$, the first derivative dJ_{BBT}/dV would not be continuous at zero bias. This would be undesirable for a compact model formulation because it could lead to convergence problems.

A compact model implementation of both a vanishing band-to-band tunneling current in forward bias and a continuous derivative at the transition from reverse to

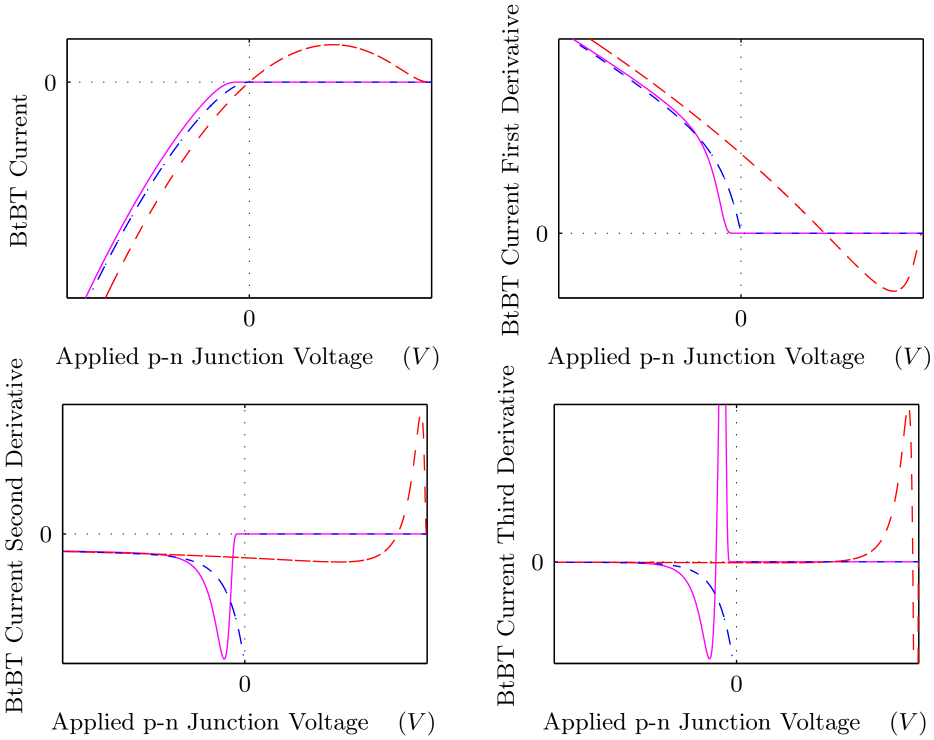

Figure 5.3 – Zeroth, first, second and third derivative of several models of the band-to-band tunneling current with respect to applied junction voltage V, all plotted as a function of V. The horizontal axes have the same scale, while the scale of the vertical axis is adjusted so as to capture details. A model (i) where the state occupancy factor D (5.2) is approximated with its leading term $-qV$ is represented by the dashed lines, a model (ii) based on a straightforward implementation of expression (5.2) is represented by a dash-dotted lines, while my actual model (iii) implementation featuring a smooth transition at zero bias is represented by solid lines. For models (ii) and (iii), the tunneling current J_{BBT} is implemented as identically vanishing in forward bias.

forward bias, can be established by adopting expression (5.2) to represent the factor D. The resulting model is represented by the dash-dotted curve in Figure 5.3. In order to have also the higher order derivatives of the model continuous, some further straightforward, but *ad hoc*, manipulation of the equations is required. The technical details of this are discussed in the Appendix B. and the resulting model is represented by the solid curves in Figure 5.3.

In practice, the band-to-band tunneling current around zero bias is usually masked by trap-assisted tunneling (TAT) and Shockley-Read-Hall (SRH) recombination currents. Against this background, further physical details of the D factor, such as discussed in literature [91] would not be relevant for practical implementations of semiconductor compact models, and expression (5.2) for D is a suitable one for compact model applications in integrated circuit design.

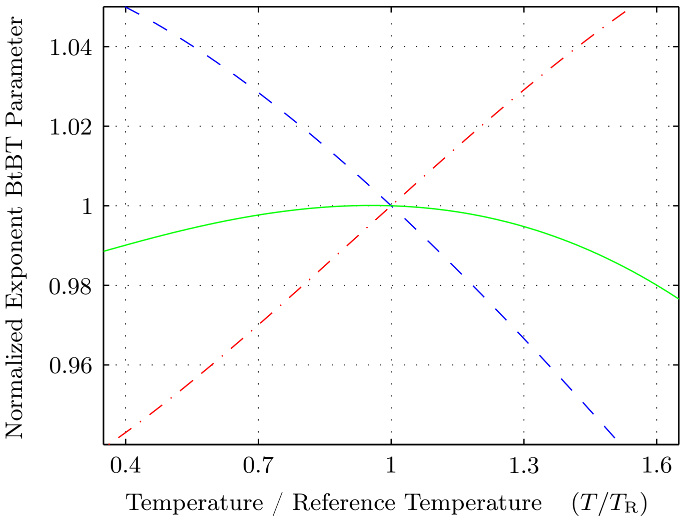

Figure 5.4 – Solid line: normalized value $\mathrm{A_{zT}/A_z}$ of the band-to-band tunneling parameter $\mathrm{A_{zT}}$ (5.11) as a function of absolute temperature normalized to a reference temperature of 300 K; $\mathrm{A_{zT}}$ is the parameter of the exponential function in expression (5.8). Dashed line: idem when temperature dependence of the electric field in band-to-band tunneling current is neglected. Dash-dotted line: idem when only temperature dependence of the band gap in BtBT current is neglected. The two effects counteract, but do not fully compensate one another – a 2 % offset at 480 K is observed.

5.2.3 Temperature dependence and scaling

The temperature dependence of the tunneling current has two physical origins. Firstly, the maximum electric field in p-n junctions depends on temperature. Secondly, the band gap energy E_g depends on temperature. Both effects turn out to have comparable significance, whereas the overall temperature dependence of the tunneling current is comparatively minor. Against this background, several approaches in implementing the temperature dependence of the tunneling current have been reported in the compact modeling literature. In [55] it was observed that both effects more or less neutralize each other and hence could both be neglected in the argument of the exponential function. In contrast to that in [85], the temperature dependence of the BtBT has been entirely represented by an empirical temperature dependence of the argument of the same exponential factor. In Figure 5.4 the temperature dependence of the normalized exponent tunneling parameter $\mathrm{A_{zT}/A_z}$ (5.11) is analyzed. When the temperature dependence of the electric field is neglected (dashed line), the tunneling current decreases with increasing temperature. The temperature dependence through the electric field, while the band gap is made constant as a function of temperature, (dash-dotted line) is just the opposite. Both effects counteract but do not fully compensate one another: as is shown, the effective representative model parameter

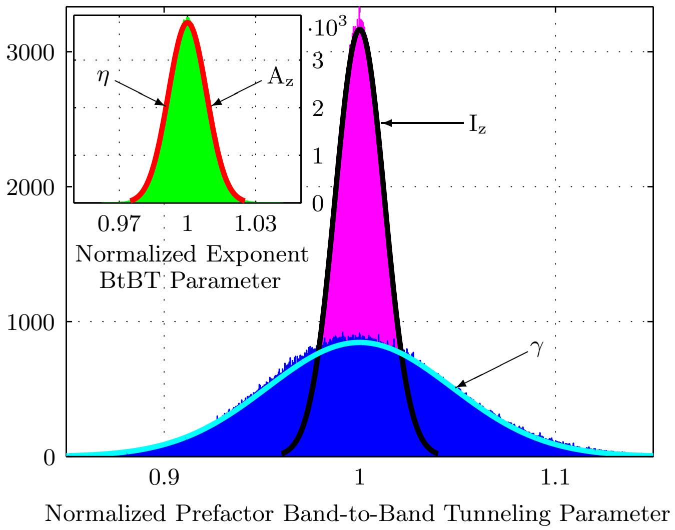

Figure 5.5 – Histograms of extracted normalized band-to-band tunneling parameter values. Parameters were extracted from one million different, numerically generated data sets. Each data set consisted of simulated $I_{\mathrm{BBT}}(V)$ data, with added numerically simulated normally distributed "stochastic measurements errors". Curves show fitted normal distribution probability density functions. The inset shows a histogram (surface) of the exponent parameters η and $\mathrm{A_z}$. Since relation (5.6) between the two parameter definitions is in fact only scaling, the normalized distributions of η and $\mathrm{A_z}$ are identical. The main plot presents corresponding results for prefactor parameters γ and $\mathrm{I_z}$, used in (5.4) and (5.8), respectively. The figure shows that the variance of the extracted values of $\mathrm{I_z}$ is much smaller than the variance associated with γ. This explicitly demonstrates that the sensitivity of a model parameter to stochastic measurement errors (noise) can depend dramatically on the definition of the parameter.

deviates up until 2% from its value at reference temperature (300 K). Based on these observations, I have chosen to retain the full physics-based temperature dependence, as implied by expressions (5.1) to (5.3), in the model formulation. The merits of this will be experimentally demonstrated in Section 5.4.

5.2.4 Parameter definition and geometrical scaling

A straightforward parametrization [85, 55] of expression (5.1) reads as

$$I_{\mathrm{BBT}} = \gamma\, E_{\mathrm{max}}\, D \exp\left(-\eta \,/\, E_{\mathrm{max}} \right) \quad , \tag{5.4}$$

where γ and η are the compact model parameters, and E_{max}, is the normalized, dimensionless maximum electric field in the p-n junction

$$E_{\mathrm{max}} = \left(1 - V \,/\, \varphi_{\mathrm{i}} \right)^{\,1-p} \quad . \tag{5.5}$$

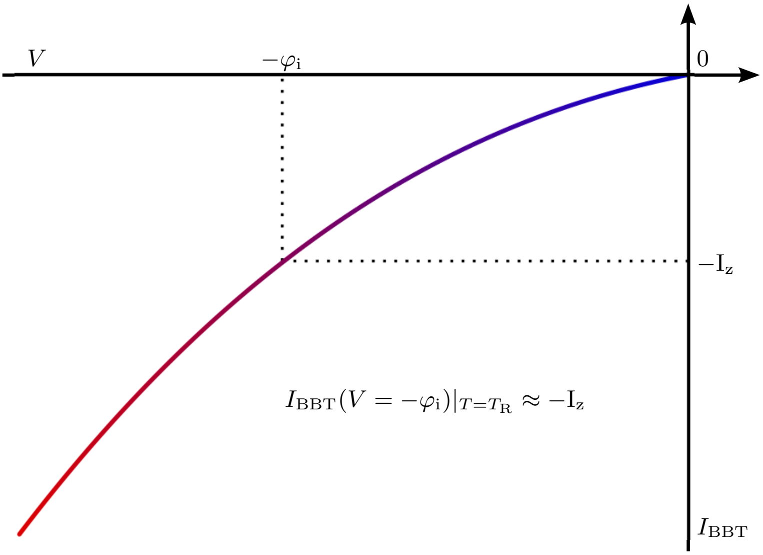

Figure 5.6 – Qualitative representation of the novel prefactor band-to-band tunneling current compact model parameter definition. When the applied p-n junction reverse bias voltage equals the built-in potential $V = -\varphi_\mathrm{i}$, the BtBT current at reference temperature approximately equals the value of the prefactor parameter $I_\mathrm{BBT} \approx -\mathrm{I_z}$. In this way the parameter value can be directly read from the measured curves.

In this last expression φ_i is the temperature dependent junction built-in potential (diffusion voltage) and p is the junction grading coefficient [85], see Appendix C.

Parametrization as in (5.4) has the disadvantage that estimated values of the parameters, as obtained by fitting of the model to measured data, are very sensitive to stochastic errors in the measured data. This can be demonstrated as follows.

Based on expressions (5.4) and (5.5), an $I_\mathrm{BBT}(V)$ characteristic is simulated for parameter values of $\gamma = 1\,\mathrm{S}$, $\eta = 10$, $\varphi_\mathrm{i} = 1\,\mathrm{V}$ and $p = 1/2$ at a hundred equidistant voltage points in the interval $V \in [-2.5, -0.5]\,\mathrm{V}$. To simulate stochastic measurement errors, artificially generated normally distributed pseudorandom numbers with zero mean and standard deviation $\mathcal{N}(0,\, 2 \times 10^{-4})$ are added to these data points. Expressions (5.4) and (5.5) are then fitted to the resulting data set, to obtain parameter values for γ and η (it is assumed that values of φ_i and p are already known without error). Normalized values of η and γ as obtained from one million (10^6) repetitions of this numerical experiment are shown in histograms in Figure 5.5. The distribution of extracted values for γ thus found is about a factor 5 wider than the distribution of η. As will be demonstrated later in Figure 5.9, as a consequence, scaling of the parameter γ over device geometry is hard fought.

The demonstrated sensitivity of the model parameter γ can be remedied by the following compact model parameter transformations

$$\mathrm{A_z} \quad = \quad 2^{\,p\,-\,1} \cdot \eta \quad , \tag{5.6}$$

Table 5.1 – Literature [90] values of the material constants in (5.12).

Semiconductor	$V_{gz}(0)$ [V]	α [V/K]	β [K]
Germanium (Ge)	0.7437	4.8×10^{-4}	235
Silicon (Si)	1.1692	4.9×10^{-4}	655
Gallium Arsenide (GaAs)	1.519	5.4×10^{-4}	204

$$\mathrm{I_z} \quad = \quad 2^{\,1-p} \cdot \exp\left(-\mathrm{A_z}\right) \cdot \gamma \quad . \tag{5.7}$$

This leads to the band-to-band tunneling current model implementation

$$I_{\mathrm{BBT}} = \mathrm{I_{zT}} \frac{E_{\mathrm{maxT}} D_{\mathrm{T}}}{2^{1-p} \varphi_{\mathrm{iT}}} \exp\left[\mathrm{A_{zT}} \left(1 - \frac{2^{1-p}}{E_{\mathrm{maxT}}}\right) - \frac{\delta}{V^2}\right] , \tag{5.8}$$

while the state occupancy factor D_{T} is implemented as

$$D_{\mathrm{T}} = -V - \frac{V_{gz}\, E_{\mathrm{maxT}}}{2^{2-p} \mathrm{A_{zT}}} \left[1 - \exp\left(\frac{2^{2-p} \mathrm{A_{zT}} V}{V_{gz}\, E_{\mathrm{maxT}}}\right)\right] \quad . \tag{5.9}$$

The parameters $\mathrm{I_z}$ and $\mathrm{A_z}$ have been introduced such that, when reverse bias equals built-in potential, $V = -\varphi_{\mathrm{i}}$, and hence $E_{\mathrm{max}} = 2^{1-p}$ according to expression (5.5), the dominant dependence on the parameter $\mathrm{A_z}$ of the current I_{BBT} through the exponential term in (5.8) is suppressed, while simultaneously $D \approx -\varphi_{\mathrm{i}}$. As a result, in the context of parameter extraction, the value of the parameter $\mathrm{I_z}$ is practically defined as the value of the (band-to-band tunneling) current observed at bias condition $V = -\varphi_{\mathrm{i}}$, as shown in the drawing of Figure 5.6.

Normalized statistical distributions of parameter values for $\mathrm{I_z}$ and $\mathrm{A_z}$, as extracted from the very same numerically generated data sets, described earlier, are shown in Figure 5.5. Since transformation (5.6) only scales parameter $\mathrm{A_z}$, the associated normalized variability remains unchanged. In contrast, the effect on the prefactor parameter $\mathrm{I_z}$ is substantial. The probability density function of the parameter $\mathrm{I_z}$ with novel definition features much lower estimated parameter variance comparative to the variance of the more straightforward parameter γ. The actual ratio between the standard deviation of γ and that of $\mathrm{I_z}$ increases monotonically with the value of $\mathrm{A_z}$.

5.3 Compact model implementation

The model of band-to-band tunneling (BtBT) current introduced above has been implemented in accordance with the discussion in Section 5.2. The contribution to the current across the p-n junction due to band-to-band tunneling effects is assumed to vanish in forward bias mode. Hence, $I_{\mathrm{BBT}} = 0$ whenever $V \geq 0$. In reverse bias mode, $V < 0$, it is modeled by the expressions (5.8) and (5.9). The reverse bias p-n junction BtBT current I_{BBT} is defined to be positive if it flows from p node region towards n node region. As will be discussed in detail in the Appendix B, to implement

Table 5.2 – Parameter Extraction Procedure Overview

Measurement	Obtained Data	Extracted Parameters
Forward Ideal Current	$I(V,T)$	$I_{\mathrm{S}}(T_{\mathrm{R}})$, $V_g(T_{\mathrm{R}})$
Depletion Capacitance	$C_{\mathrm{j}}(V,T)$	$C_{\mathrm{j0}}(T_{\mathrm{R}})$, $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$, p
BtB Tunneling Current	$I(V)\vert_{T=T_{\mathrm{R}}}$	$\mathrm{I_z}$, $\mathrm{A_z}$
BtB Tunneling Current	$I(V,T)$	$V_{g\mathrm{z}}(0)$, α, β

a smooth transition from reverse bias, to zero forward bias tunneling current at $V = 0$, in the argument of the exponent of (5.8) a term $-\delta/V^2$ is added.

As well documented in the compact modeling literature [85] as well as in Appendix C, the compact model for the maximum value of the electric field in a junction can be shared with the junction depletion capacitance compact model (together with its temperature dependence). The temperature scaling rules for the band-to-band tunneling current compact model parameters $\mathrm{A_{zT}}$ and $\mathrm{I_{zT}}$ follow straightforwardly from the physical foundation presented in detail in Section 5.2.3:

$$\mathrm{A_{zT}} = \mathrm{A_z}\left(\frac{V_{g\mathrm{z}}(T)}{V_{g\mathrm{z}}(T_{\mathrm{R}})}\right)^{3/2}\left(\frac{\varphi_{\mathrm{i}}(T_{\mathrm{R}})}{\varphi_{\mathrm{i}}(T)}\right)^{1-p} \quad , \tag{5.10}$$

$$\mathrm{I_{zT}} = \mathrm{I_z}\left(\frac{V_{g\mathrm{z}}(T_{\mathrm{R}})}{V_{g\mathrm{z}}(T)}\right)^{1/2}\left(\frac{\varphi_{\mathrm{i}}(T)}{\varphi_{\mathrm{i}}(T_{\mathrm{R}})}\right)^{2-p} \exp(\mathrm{A_z} - \mathrm{A_{zT}}) \quad . \tag{5.11}$$

The model parameter $V_{g\mathrm{z}}(T_{\mathrm{R}})$ is the band gap voltage at reference temperature T_{R}. The band gap voltage $V_g = E_g/q$ of common semiconductor materials is known to be approximately given by the well-known textbook [92] expression

$$V_{g\mathrm{z}}(T) = V_{g\mathrm{z}}(0) - \frac{\alpha \cdot T^2}{T+\beta} \quad , \tag{5.12}$$

in which T is the absolute temperature (the Kelvin scale). The values of the material parameters $V_{g\mathrm{z}}(0)$, α and β are tabulated for the most common semiconductor materials in Table 5.1. As it shall be demonstrated on a typical industrial SiGe BiCMOS SOI process in Section 5.4, these literature values represent practical estimate values of the model parameters. In the actual model implementation, the quantity $V_{g\mathrm{z}}(0)$ is a model variable, the value of which then follows from expression (5.12), as evaluated at $T = T_{\mathrm{R}}$. The scaled band gap voltage $V_{g\mathrm{z}}(T)$ at device simulation temperature T, subsequently follows from relation (5.12), as well.

5.4 Parameter extraction and model verification

A full procedure for extraction of all relevant parameters is outlined in Table 5.2. Since the band-to-band tunneling current is modeled in terms of the electric field in the p-n junction's space charge region and since the electric field is also a key quantity in the

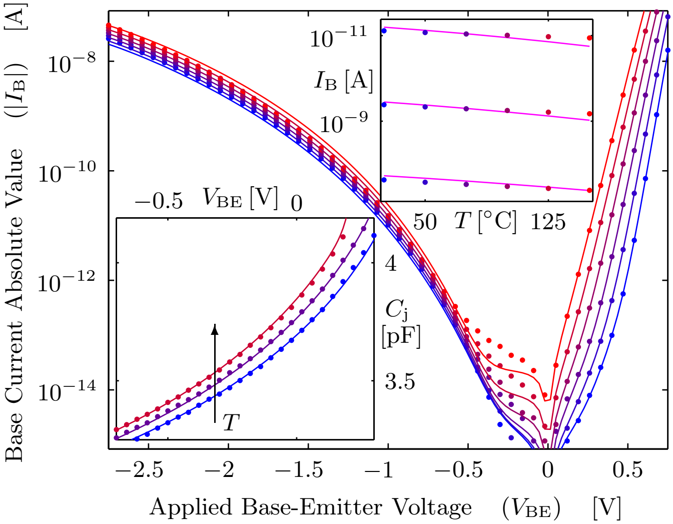

Figure 5.7 – Measured base current (markers) of a typical industrial SiGe BiCMOS HBT as a function of base-emitter voltage V_{BE} at six different temperatures $T \in \{25, 50, 75, 100, 125, 150\}$°C and for base-collector voltage $V_{\mathrm{BC}} = 0\,\mathrm{V}$. The top-right inset shows the same data (symbols) as a function of temperature for three base-emitter voltages $V_{\mathrm{BE}} \in \{-0.96, -1.6, -2.7\}\,\mathrm{V}$. The lower left inset in shows measured (markers) and simulated (lines) base-emitter capacitance C_{j} as a function of applied junction voltage at three different temperatures $T \in \{25, 75, 125\}$°C, from which depletion capacitance parameters were extracted. Figure shows this data together with simulations (lines) based on a simplified version of our BtBT model, in which temperature scaling of the exponent parameter $\mathrm{A_{zT}}$ has been switched off. The reduced version of the model shows a systematic deviation in this respect (top inset).

modeling of depletion capacitance, see Appendix C, the junction's depletion capacitance parameters $C_{\mathrm{j0}}(T_{\mathrm{R}})$, $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$ and p are extracted first, prior to the extraction of the dedicated BtBT current parameters. Here, $C_{\mathrm{j0}}(T_{\mathrm{R}})$ denotes the depletion capacitance at zero bias, $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$ the junction diffusion voltage (built-in potential), both at reference temperature, and p the grading coefficient. These parameters were extracted by a simultaneous [85] fit of simulated capacitance to measured $C_{\mathrm{j}}(V,T)$ data, that is, capacitance as a function of bias at several temperatures, $T \in \{25, 75, 125\}$°C. Because this leans on the temperature scaling rules of C_{j0} and φ_{i}, which in turn also involve the band gap voltage V_g, the band gap voltage at reference temperature $V_g(T_{\mathrm{R}})$ is extracted in an earlier extraction stage. It is done so by a fit of the ideal forward p-n junction current $I(V,T)$ characteristics over several (six) temperatures $T \in \{25, 50, 75, 100, 125, 150\}$°C, having the band gap voltage $V_g(T_{\mathrm{R}})$ and junction ideal reverse bias saturation current $I_{\mathrm{S}}(T_{\mathrm{R}})$ as optimization parameters.

Given extracted values of the parameters mentioned above, extraction of the ded-

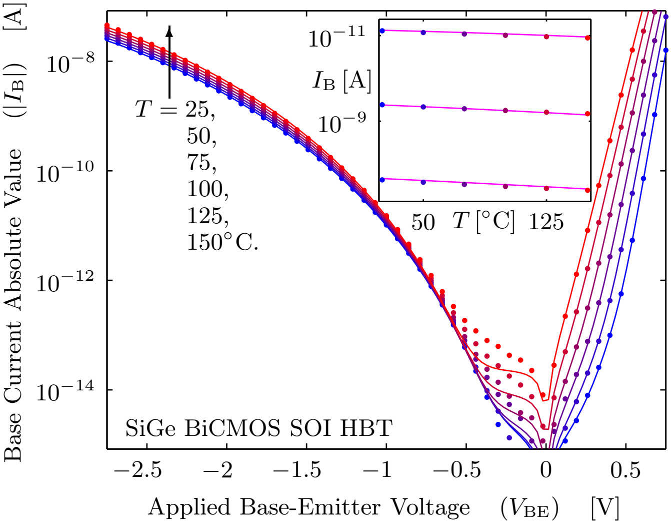

Figure 5.8 – Measured base current (markers) of a typical industrial SiGe BiCMOS HBT as a function of base-emitter voltage V_BE at six different temperatures $T \in \{25,\ 50,\ 75,\ 100,\ 125,\ 150\}$°C and for base-collector voltage $V_\mathrm{BC} = 0\,\mathrm{V}$. The top-right inset shows the same data (symbols) as a function of temperature for three base-emitter voltages $V_\mathrm{BE} \in \{-0.96,\ -1.6,\ -2.7\}\,\mathrm{V}$. The figure shows the same data together with simulations (lines) based on a full version of our BtBT model, which includes the full physical temperature scaling of both tunneling parameters. The full model, simulates the temperature dependence of the data within measurement accuracy.

icated BtBT parameters is straightforward. The relevant measurement setup for the isolated diode is trivial. For bipolar transistors, measurements can be taken in the same setup as the forward Gummel measurements, keeping the base-collector junction at zero bias $V_\mathrm{BC} = 0$ and sweeping the base-emitter bias V_BE into the reverse regime.

The dedicated band-to-band tunneling parameters, $\mathrm{I_z}$ and $\mathrm{A_z}$, are extracted simultaneously by fitting the simulated p-n junction reverse current to the measured one at reference temperature, focusing on the band-to-band tunneling regime. This regime is best recognized in a plot that shows the observed reverse (base) current as a function of applied bias voltage V for several temperatures T. In the low bias regime, the current may be dominated by trap-assisted tunneling effects [85], that can be identified by their strong temperature dependence. At higher bias conditions, the band-to-band tunneling current becomes dominant. This regime can be easily identified by its weak temperature dependence, as observed in Figures 5.7 and 5.8.

The developed model of BtBT current was implemented in the Mextram 504, physics-based standard compact model for vertical bipolar transistors. To verify the model at reference temperature as well as its temperature scaling rules, it was applied to model the characteristics of an NPN-type bipolar transistor manufactured in an

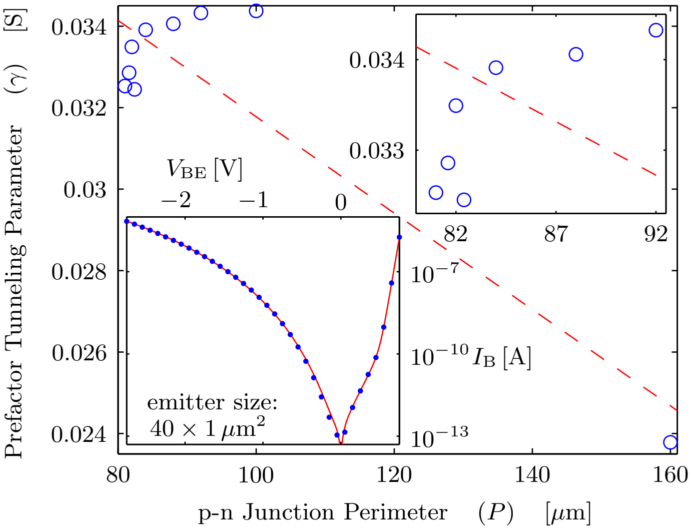

Figure 5.9 – Prefactor parameter of two different band-to-band tunneling current models as a function of emitter(-base junction) perimeter for a series of Dimes bipolar transistors with emitter lengths of 40 μm and widths of 0.5, 0.8, 1, 1.2, 2, 4, 6, 10 and 40 μm. Shown are the compact model parameter values as extracted for individual devices (markers) and the first order polynomial fit (dashed line). Top right inset shows details of the main plot. Shown results are obtained on the basis of the reduced version of the BtBT model according to expression (5.4). This model basically follows from straightforward parametrization of classical textbook expressions for BtBT current. In lower left inset is the typical result for measured data (markers) and model simulation (curve) of the base terminal current as a function of applied base-emitter bias voltage for an individual device. Although modeling results at the level of individual devices are excellent, the geometrical scalability of the extracted parameter γ is pretty poor.

industrial SiGe BiCMOS SOI technology. Figures 5.7 and 5.8 present data taken on such a bipolar device that has one emitter contact with emitter size $0.4 \times 0.8\,\mu\text{m}^2$. Relevant parameters of Mextram were extracted following the strategy outlined above.

To verify the significance of the full physical temperature scaling rules (5.10) and (5.11), the full model is compared to a reduced version of it, which has simplified temperature scaling rules, as follows. Figure 5.8 presents simulations (lines) of the full model together with measured data (markers). Figure 5.7 represents the same measured data together with best fits of a reduced version of the model, in which the temperature dependent parameter $\mathrm{A_{zT}}$ was replaced by a parameter of constant value $\mathrm{A_z}$. This approximation is suggested [55] by the fact that the temperature dependencies of $\mathrm{A_{zT}}$ as induced by temperature dependence of the built-in voltage φ_i and that of the band gap approximately compensate one another. For both figures the value of the band gap $V_{gz}(T_\mathrm{R})$ parameter was set equal to the value of the junction band gap

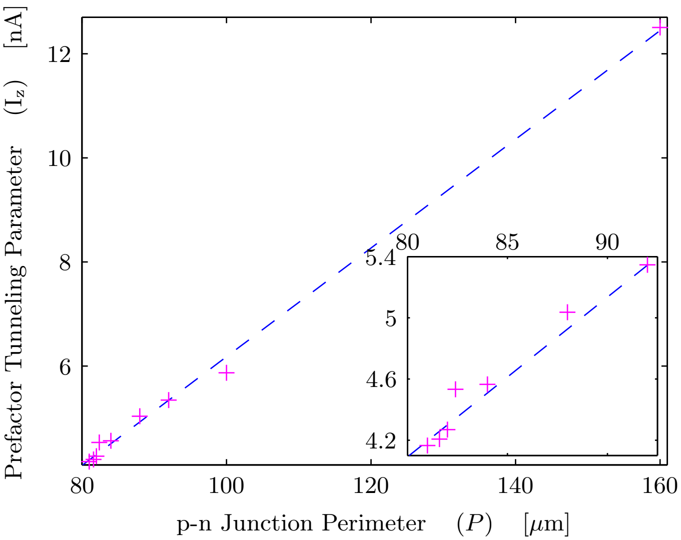

Figure 5.10 – Prefactor parameter of two different band-to-band tunneling current models as a function of emitter(-base junction) perimeter for a series of bipolar transistors with emitter lengths of 40 μm and widths of 0.5, 0.8, 1, 1.2, 2, 4, 6, 10 and 40 μm, the same ones as in Figure 5.9. Shown are the model parameter values as extracted for individual devices (markers) and the first order polynomial fit (dashed line). Lower right inset shows details of the main plot. Results are the ones obtained on basis of the model (5.8), which has an innovative parametrization. Much better geometrical scalability of the extracted parameter values correspond to much lower sensitivity of the extracted parameter values to stochastic measurement errors (see also Figure 5.5).

voltage, V_g, as extracted from the temperature dependence of the ideal forward base current. For the parameter β, the literature value for silicon (Si) was used. The simulated base current in the BtBT regime was then optimized with respect to measured characteristics for six temperatures, with only α as a free parameter. The full model gives excellent fits, for $\alpha = 9.6 \times 10^{-4}$ V/K, which is comparable to the known value for intrinsic silicon, as quoted in Table 5.1. The fit of the reduced model, in contrast, shows a significant systematic error in temperature dependence for the nonphysical, but the best fit value of $\alpha = -155 \times 10^{-4}$ V/K.

Geometrical scalability of the presented model was verified on nine NPN-type bipolar junction transistors of our in-house DIMES04 technology. The geometric scaling of parameter I$_z$ is shown in Figure 5.10. For comparison, Figure 5.9 shows corresponding result obtained on basis of model based on parametrization (5.4). Figures 5.9 and 5.10 and shows that the alternative parametrization, of Figure 5.10, remedies the scaling problems that are associated with more straightforward parametrization; scaling results based on the latter are shown in Figure 5.9. Figure 5.10 shows that the parameter I$_z$ is proportional to the perimeter of the emitter(-base junction). This is in accordance with the fact that the devices are mid-speed mid-power ones, and do not possess extremely high doping in the base and emitter. In such devices, the electric field is strongest at the edges of the junction, and so is the tunneling.

5.5 Discussion and conclusion

Discussion

Presented in this chapter is the compact model of band-to-band tunneling current in p-n junctions. However, in order to have a full prediction of the junction leakage, as it may occur in modern heterojunction bipolar or deeply scaled CMOS field-effect transistors, in one single compact model, one has to add components of Shockley-Read-Hall (SRH) current and trap-assisted tunneling (TAT) current that dominate the low bias region (and become more pronounced on higher temperatures), as well as the avalanche multiplication current that ultimately limits the reverse bias voltage of p-n junctions. The model described in [85] seems to be appropriate for inclusion of all of the mentioned terms. As stated earlier, this model also includes modeling of band-to-band tunneling currents. Nevertheless, based on everything shown in this chapter, I believe that the model proposed in this chapter advances it in several very important segments. Therefore, combination of these two compact models seems as a logical step toward a complete physics-based junction leakage modeling solution.

Conclusion

A novel physics-based compact model of band-to-band tunneling (BtBT) current is presented. Band-to-band tunneling may occur in highly doped, reverse biased, p-n junctions of MOS field effect, as well as in (heterojunction) bipolar transistors. The model implements an identically vanishing tunneling current in the forward bias regime. This is in accordance with band-to-band tunneling theory and attractive in terms of computational load. Based on incorporation of a physical model of quantum states occupation effects, the model features a smooth transition, at zero bias, from the reverse bias to zero forward bias regime.

The model includes fully physics-based temperature scaling rules that take into account temperature dependencies of both the band gap and the built-in junction electric field. It has been experimentally verified on state of the art industrial SiGe HBTs, that in order to have faithful modeling results, both effects indeed have to be explicitly taken into account.

The presented BtBT model introduces a nonstraightforward parameter definition that decreases sensitivity of extracted parameter values to stochastic errors found in measured data. This results in significantly enhanced geometry scaling abilities of the model. A parameter extraction procedure is discussed and applied to measurements taken on industrial SiGe BiCMOS SOI devices and on in-house Si devices.

Chapter 6

Extraction of p-n junction compact model parameters

Two general classes of strategies for the extraction of compact model electrical and temperature scaling parameters are identified and compared in this chapter. Applying these to extract parameters of the p-n junction depletion capacitance (themselves very important because the junction electric field model depends on them) and ideal diode current, it is shown that the reproducibility of the estimated parameter values can strongly depend on the extraction strategy applied in the nonlinear regression procedure. In addition, an approach to assess statistical properties of parameter extraction strategies and demonstrate the merits of such assessments is presented.

6.1 Introduction

Semiconductor device compact modeling and parameter extraction are key concepts in Electronic Design Automation (EDA). Advancements in semiconductor technology are characterized by, for example, downscaling or introduction of advanced heterojunction technologies. To follow these developments compact models for diodes [85], field-effect transistors (FETs) [93, 94, 95] and bipolar junction transistors (BJTs) or heterojunction bipolar transistors (HBTs) [47, 49, 50] have evolved in the direction of greater complexity in order to capture subtle physical phenomena. The increasing complexity of compact semiconductor device models has made the extraction of model parameters a considerable task. The implied investments will pay only if the extracted parameter values are of good quality, for the simple reason that the predictive value of compact model based circuit simulations strongly depends on the quality of the set of parameter values that is used as the basis for them.

Since modern electronic circuit designs, such as mobile applications and car systems, have to meet explicit specifications in terms of exposure to ambient temperatures, compact models have to take into account temperature dependencies of the electrical characteristics of devices. Modeling of temperature dependencies have only become more significant with the introduction of substrate isolation techniques, which have introduced significant thermal resistances and hence increased the significance of self-heating effects. Against this background, naturally, consideration of temperature dependencies of characteristics and of model parameters associated with them will be an integral part of compact model parameter extraction efforts.

6.1.1 Compact model structure and extraction strategies

Quite in general, in compact models one can distinguish, firstly, a basic set of equations with associated electrical parameters that capture dependencies of quantities on electrical bias conditions and, secondly, a supplementary set of equations, with associated parameters that address the temperature dependencies.

Compact models are commonly formulated in a hierarchical way, such that all electrical characteristics can be evaluated at a specified reference temperature $T = T_{\mathrm{R}}$ on basis of the electrical set of parameters alone. More precisely, models are formulated commonly in such a manner, that the characteristics at the reference temperature are invariant under changes of the values of the temperature scaling parameters. In this sense, characteristics at alternative temperatures can be conceived as constructed by scaling of the corresponding characteristics at reference temperature.

This structure of the compact models suggests the possibility of applying a conformal structure in parameter extraction strategy. Indeed, one could in a first stage extract the first set of electrical parameters to capture measured characteristics at reference temperature $T = T_{\mathrm{R}}$. At that stage, the value of the dedicated temperature scaling parameters is irrelevant and outside the scope of the effort. If such an approach turns out to be successful in the first phase, one may even define a sequence of temperature values to be the reference temperature for parameter extraction in a sequence of corresponding subsequent extractions of the electrical parameters. This would yield the values of the electrical parameters as a function ($P_i(T)$, say) of temperature. In a final stage then, the temperature scaling rules of the electrical parameters P_i could then be fitted [96] to the $P_i(T)$ dependencies thus observed, so as to yield the values of the temperature scaling parameters. This first approach to the extraction of the hierarchy of parameters can also be followed in a more direct way by first fitting simulated characteristics to measured data as a function of bias (V, say) $f\,(V, T_{\mathrm{R}})$ at reference temperature, having the relevant electrical parameters as variables to be optimized. Secondly, one would fit simulated characteristics to measured data over both bias and several alternative temperatures $f\,(V, T)$, having exclusively the relevant temperature scaling parameters as variables.

The strategies as outlined above are deliberately supported and even suggested by the usual structure of compact models. They have in common the advantage of minimizing the number of variable parameters to be optimized in each stage of the extraction effort. In the present chapter such strategies shall be denoted as class I strategies. This class is characterized by the fact that data over different temperatures is used exclusively to extract temperature scaling model parameters.

Against this background, one may consider a second, apparently more trivial strategy of parameter extraction. This second strategy (II) consist of straightforward fitting of simulated characteristics to measured data at all relevant temperatures $f\,(V,\,T)$ at once, having all the relevant electrical and temperature scaling parameters as variables to be optimized. This strategy class as an output gives both electrical and temperature scaling parameters that are obtained from complete data set.

6.1.2 Quality of parameter sets obtained by extraction

Now that two classes, denoted by I and II above, of parameter extraction strategies have been identified, and given the observation in the introductory part of Section 6.1 that the quality of extracted parameter values is crucial for application of compact models, the question arises whether or not the quality of the obtained parameter values depends on the chosen strategy.

In the present chapter, it shall be considered that a parameter extraction strategy yields parameter sets of "high quality" if the parameter values obtained are well-reproducible over the separate extraction runs. A suitable underlying question is: if the measurement and parameter efforts as a whole would be repeated, so that effectively the parameters would be reextracted on an experimentally reproduced version of the set of measured data, would then a reasonably similar set of parameter values be obtained? In the present chapter, the similarity of the parameter values thus obtained shall be quantified by the parameter values variance, or standard deviation of their stochastic distribution. This stochastic distribution of the extracted parameter values is a direct consequence of the stochastic errors that will inevitably be present in the measured data that form the starting point of the extraction.

In the present chapter it shall be explicitly demonstrated, that the choice for either the first (I) or the second (II) parameter extraction strategy as outlined above, can have a vast impact on the reproducibility of the parameter values obtained.

It shall be done so by means of numerical experiments designed to mimic repeated parameter extraction on a typical industrial heterojunction focus will be drawn on the selected but very practical [97, 98, 99] case of extraction of compact model parameters of the p-n junction depletion capacitance, themselves very important due to their essential role in modeling p-n junction's electric field. It shall be shown that the reproducibility of the parameters of depletion capacitance is significantly better, in the second (II), more straightforward parameter extraction strategy. Specifics on how maximum electric field, depletion layer width or depletion region charge are expressed in terms of junction depletion capacitance are deferred to Appendix C.

6.2 p-n junction depletion capacitance model

The junction depletion capacitance C_j as a function of applied voltage V at absolute temperature T is given [100] by the well-known expression

$$C_\mathrm{j}(V,T) = \frac{C_\mathrm{j0}(T)}{(1 - V/\varphi_\mathrm{i}(T))^p} \quad , \tag{6.1}$$

where both the zero bias junction depletion capacitance C_j0 and the built-in potential (also know as diffusion voltage) φ_i depend on temperature. The junction grading coefficient p, which depends on the doping profile, see Appendix C, is temperature independent. The temperature dependence of the junction zero bias depletion capacitance $C_{\mathrm{j0}}\left(T\right)$ is governed by the built-in voltage in form of

$$C_{\mathrm{j0}}\left(T\right) = C_{\mathrm{j0}}\left(T_{\mathrm{R}}\right)\left(\frac{\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)}{\varphi_{\mathrm{i}}(T)}\right)^{p} \quad , \tag{6.2}$$

in which $C_{\mathrm{j0}}\left(T_{\mathrm{R}}\right)$ and $\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)$ in turn are the zero bias junction depletion capacitance and built-in voltage, respectively, both at reference temperature. The temperature dependence of the diffusion potential φ_{i} can be expressed as

$$\varphi_{\mathrm{i}}(T) = \varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)\frac{T}{T_{\mathrm{R}}} - 3V_{\mathrm{T}}\ln\frac{T}{T_{\mathrm{R}}} + V_g\left(1 - \frac{T}{T_{\mathrm{R}}}\right) \quad . \tag{6.3}$$

The quantities T and T_{R} are device temperature and the reference temperature for parameter extraction in Kelvins (in the Kelvin scale), $V_{\mathrm{T}} = k_{\mathrm{B}}T/q$ is the thermal voltage at device temperature (k_{B} is Boltzmann's constant, q the elementary charge) and $V_g = E_g/q$ is the band gap voltage (i.e., band gap E_g divided by the elementary charge), which represents the temperature scaling parameter in this case. Without loss of chapter's generality, in equation (6.3), the temperature dependence of the band gap is neglected for to simplicity, although it could be easily included through Varshni's [92] empirical expression, as done by (5.12) in previous chapter.

The system of equations (6.1) to (6.3) shows a hierarchical structure formed by an electrical (6.1) model equation and temperature scaling rules (equations (6.2) and (6.3)) as discussed in Section 6.1.1. To simulate the capacitance characteristic at reference temperature $C_{\mathrm{j}}(V, T_{\mathrm{R}})$, the values of the basic parameters $C_{\mathrm{j0}}\left(T_{\mathrm{R}}\right)$, $\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)$ and p would suffice. Hence, in that case basically only equation (6.1) needs to be evaluated, whereas equations (6.2) and (6.3) reduce to trivialities. Only when one desires to simulate temperature dependence of the capacitance, exists a need for the temperature equations (6.2) and (6.3) and the additional temperature scaling parameter, the band gap (voltage) V_g, to be taken into account.

6.3 On statistical errors in measurements

To establish a first direct comparison of the two p-n junction depletion capacitance parameter extraction strategies I and II, identified in Section 6.1.1, both were applied on data obtained measuring (with an impedance analyzer) the capacitance of a typical industrial p-n heterojunction that might be used as the base-emitter junction of a bipolar transistor. The on-wafer capacitance measurements using a temperature chuck were taken on five different temperatures $T \in \{20, 50, 75, 100, 125\}$°C, while the anode-cathode voltage V was swept from $-0.5\,\mathrm{V}$ until $0.2\,\mathrm{V}$ with steps of $0.01\,\mathrm{V}$. A reference temperature for parameter extraction was chosen to be $T_{\mathrm{R}} = 20$°C. The band gap $E_g = 1.144\,\mathrm{eV}$ was independently extracted from the temperature dependence of the measured ideal forward p-n junction current. Nonlinear regression analysis, following either one of the two strategies yielded parameters as presented in

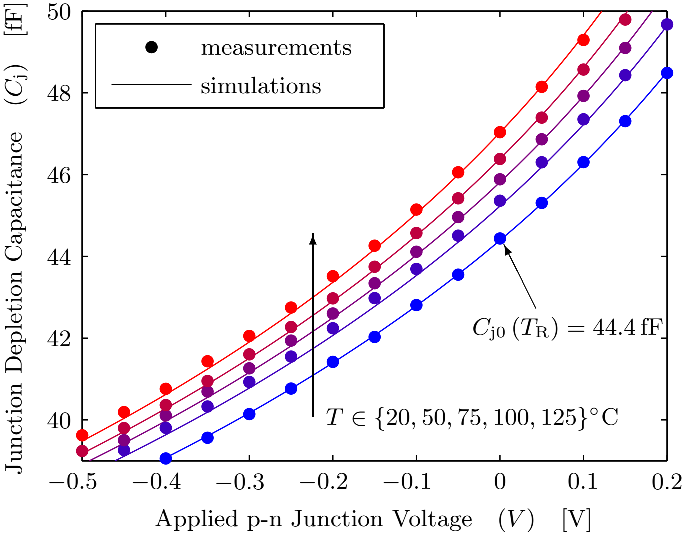

Figure 6.1 – Measured (markers) and simulated (lines) values of the p-n heterojunction depletion capacitance $C_{\mathrm{j}}\,(V,\,T)$ as a function of applied anode-cathode bias voltage V, and at five different temperatures $T \in \{20,\ 50,\ 75,\ 100,\ 125\}$°C. The zero bias depletion capacitance at reference temperature, $C_{\mathrm{j0}}\,(T_{\mathrm{R}})$, is indicated on the plot.

Table 6.1. Measured data and optimized simulated characteristics resulting from the second approach are presented in Figure 6.1.

A first observation from Table 6.1 is that, while both extraction strategies yield identical values for the zero bias depletion capacitance $C_{\mathrm{j0}}\,(T_{\mathrm{R}})$ at reference temperature, the obtained values for $\varphi_{\mathrm{i}}\,(T_{\mathrm{R}})$ and p are significantly different. This provides a further motivation to investigate the respective variances of the obtained parameter values for both estimation strategies. It will be done so in the next section by numerical simulation of the full measurement and the corresponding extraction effort.

In these simulations, measurement errors will be mimicked, numerically. Indeed, close inspection of Figure 6.1 reveals deviations between measured and simulated values. The nature of these deviations will be partly systematic – for example due to the fact that the simple one-dimensional depletion capacitance model (6.1) to (6.3) neglects higher dimensional geometrical effects – and partly stochastic. A thorough analysis of the nature of the deviations is beyond the scope of the present chapter. Instead, I have chosen the pragmatic approach to derive a practical order of magnitude of the errors from the results presented in Figure 6.1, and use that as input for a numerical study of the propagation of stochastic errors in the measured data toward extracted parameter values. Based on the data presented in Figure 6.1, it can be estimated that a normally distributed stochastic variable with a relative standard deviation of 0.1 % to 1 % would be representative for errors that occur in modeling of depletion capacitances in an industrial context.

Table 6.1 – Comparison of estimated (by nonlinear regression fit) junction depletion capacitance parameter values, following two different parameter extraction strategies.

Extraction Methodology	$C_{\mathrm{j0}}\,(T_{\mathrm{R}})$	$\varphi_{\mathrm{i}}\,(T_{\mathrm{R}})$	p
I : only reference temperature used	44.4 fF	0.99 V	0.43
II: five different temperatures used	44.4 fF	0.91 V	0.39

6.4 Numerical simulation of error propagation

The constructed numerical simulation of the total effort of experimental data collection and model parameter estimation of a given p-n junction depletion capacitance consists of the following steps. In a first step, a measured depletion capacitance data set is mimicked, by numerically evaluating

$$\tilde{C}_{\mathrm{j}}\,(V,\,T) = C_{\mathrm{j}}\,(V,\,T) + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}\,(0,\,\sigma_{\mathrm{n}}) \quad . \tag{6.4}$$

In this expression, $\tilde{C}_{\mathrm{j}}\,(V,\,T)$ represents the mimicked measured value of the depletion capacitance at the bias and temperature condition $(V,\,T)$, $C_{\mathrm{j}}(V,\,T)$ represents the value of the capacitance as predicted by expressions (6.1) to (6.3) for a fixed given set of parameter values, whereas ϵ is a normally distributed pseudorandom variable with zero mean and σ_{n}^2 variance $\mathcal{N}\,(0,\,\sigma_{\mathrm{n}})$; ϵ represents the stochastic measurement error in each of the individual measurements. In a second step, the parameters are extracted following either one of the two parameter extraction strategies. The synthetic data set $\tilde{C}_{\mathrm{j}}\,(V,\,T)$ from which the parameters are then estimated is identical for both of the extraction strategies. In this way the statistical properties of both estimation methods can be fairly compared.

Such a simulation, in all its simplicity, has two advantages over doing the exercise based on real (measured) data. Firstly, the exact values of the parameters are known, so that one can easily compare the extracted parameter values to the parameter values that are used as input for the evaluation of expression (6.4). Secondly, the whole procedure can be readily repeated many times so that statistical information can be readily generated and statistical significance quickly achieved.

6.4.1 Numerical synthesis of data

Data sets of depletion capacitance values, $\tilde{C}_{\mathrm{j}}\,(V,\,T)$, have been generated as outlined in the previous subsection at fifty-one applied bias voltages within the range $V \in [-0.7,\,0.3]$ Volts, and seven temperature steps of 25 K from the interval [275, 425] Kelvins. Without loss of generality, the zero bias depletion capacitance at reference temperature was normalized, $C_{\mathrm{j0}}\,(T_{\mathrm{R}}) = 1$, while the built-in potential at reference temperature and the grading coefficient were chosen to match certain physical values $\varphi_{\mathrm{i}}\,(T_{\mathrm{R}}) = 0.91\,\mathrm{V}$ and $p = 0.46$, respectively. Based on practise in p-n junction modeling [85], in which band gap voltages are extracted on the basis of ideal DC forward bias p-n junction current temperature dependency, the band gap voltage is assumed to be known. It has been set to $V_g = 1.11\,\mathrm{V}$. The procedure for extraction

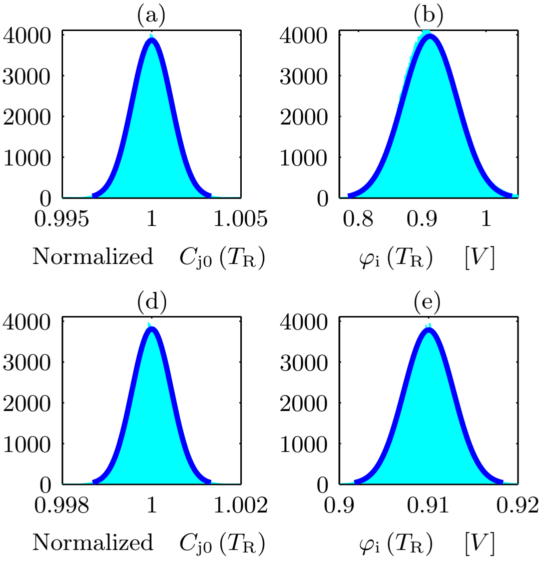

Figure 6.2 – Histograms (light surfaces), with corresponding fitted normal distributions (curves), of estimated p-n junction depletion capacitance parameters $C_{\mathrm{j0}}\left(T_{\mathrm{R}}\right)$, $\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)$ and p as follow from parameter extraction of electrical parameters, exclusively on basis of the electrical model at reference temperature (the first strategy – top row) and from global fitting of the same parameters on the basis of a full bias and temperature dependent model on data corresponding to measurements at seven different temperatures (the second strategy – bottom row). Histograms result from 10^6 independently simulated repetitions of data measurement followed by parameter extraction.

the band gap will be explained in Section 6.5 in detail. Measurement errors ϵ were represented by normally distributed pseudorandom numbers having zero mean and a relative standard deviation of 0.5 %, which given the normalized value of the depletion capacitance comes down to a stochastic relative measurement error of 0.5 %. The value of the measurement error standard deviation was chosen to be 0.5 % (between 0.1 % and 1 %) based on the discussion in Section 6.3. The reference temperature for parameter extraction was chosen to be $T = T_{\mathrm{R}} = 300\,\mathrm{K}$. The full cycle of data generation and parameter estimation was repeated 10^6 times for each of the two evaluated parameter extraction strategies. To ensure fairness of comparison, the two parameter extraction methodologies were applied to one and the same synthetic data set.

6.4.2 Results of the experiment

Figures 6.2(a) to 6.2(f) present results for numerical experiments as outlined in the introductory part of Section 6.4, carried out for two parameter extraction strategies

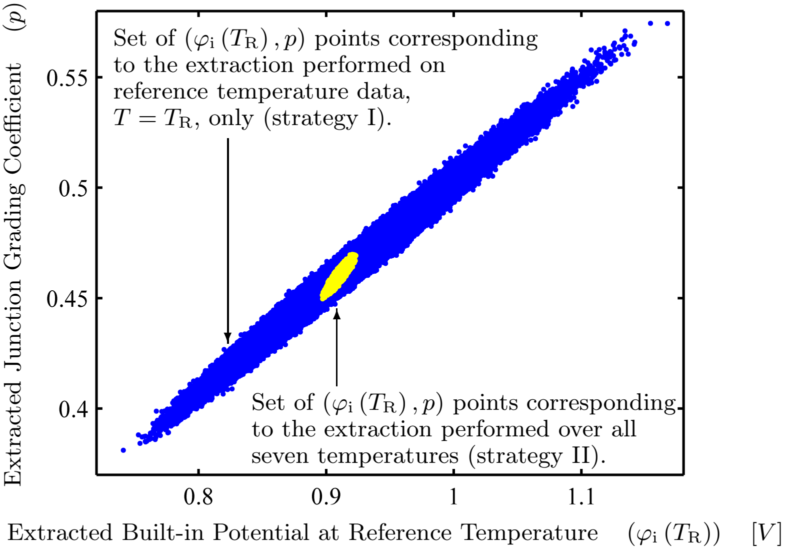

Figure 6.3 – Distribution of one million extracted pairs $(\varphi_\mathrm{i}(T_\mathrm{R}), p)$ of the junction built-in potential at reference temperature $\varphi_\mathrm{i}(T_\mathrm{R})$ and the junction grading coefficient p. The light distribution inside the dark one represents points estimated on seven temperatures, that is, utilizing the second strategy, clearly showing that such a parameter extraction strategy results in a much lower variance in extracted parameter values. High mutual parameter correlation is clearly present in both strategies, but is higher in the first one.

I and II as identified in Subsection 6.1.1. In the first strategy, represented by the upper row of figures, figures 6.2(a) to 6.2(c), the parameters have been extracted by basically fitting the electrical model, equation (6.1), to data at reference temperature only. In the second strategy, represented by the lower row of figures, figures 6.2(d) to 6.2(f), the parameters have been extracted by fitting the full bias and temperature dependent capacitance model, consisting of equations (6.1) to (6.3), to capacitance data generated over both a voltage range and at seven different temperatures, as specified in Section 6.4.1. In Figures 6.2(a) to 6.2(f), histograms of extracted parameter values are shown together with normal distribution probability density functions fitted to them. The histograms show that the parameter $C_{\mathrm{j}0}(T_\mathrm{R})$ in both cases are extracted with relatively low and mutually comparable variance. In contrast, the results for parameters $\varphi_\mathrm{i}(T_\mathrm{R})$ and p are significantly different for both extraction strategies, the most eye-catching difference being that the standard deviation σ of the obtained parameter value distribution is lower by a full order of magnitude in case of the second (II) parameter extraction strategy. This difference is also clearly illustrated in Figure 6.3, in which all pairs of extracted values $(\varphi_\mathrm{i}(T_\mathrm{R}), p)$, as obtained from the experiment have been plotted. Note that the order of magnitude of the variance as encountered by strategy I meets the differences in extracted parameter values by the

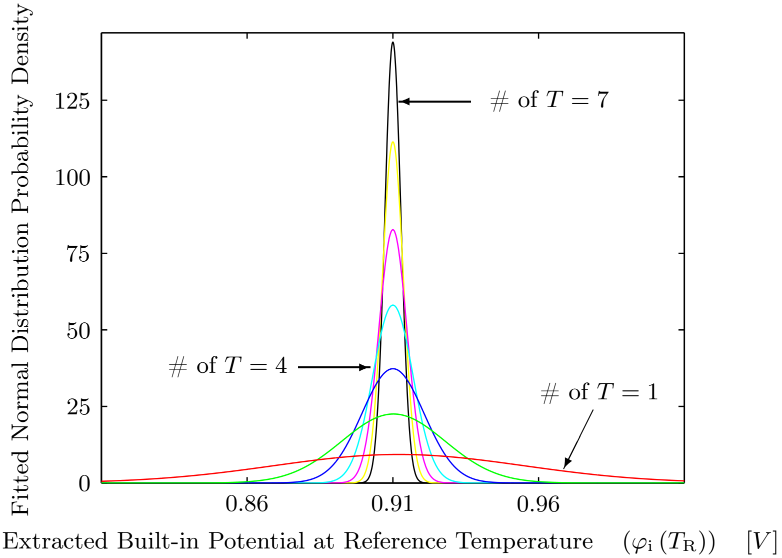

Figure 6.4 – Normal distribution probability density functions that were fitted to the histograms of estimated p-n junction built-in potential $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$ at reference temperature with different numbers of temperatures used within the extraction process. Expectably, variances (standard deviations) of statistical distributions of extracted parameter values decrease as the number of used temperatures is increased.

two methods as represented in Table 6.1. Also note that Figure 6.3 shows that the statistical correlation between the two parameters is strong in both extraction cases. Indeed, calculated correlation coefficients amount to $\operatorname{corr}(\varphi_{\mathrm{i}}(T_{\mathrm{R}}), p) = 0.992$ for the first strategy and $\operatorname{corr}(\varphi_{\mathrm{i}}(T_{\mathrm{R}}), p) = 0.903$ for the second strategy.

A second observation that can be made is that histograms of $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$ and p as extracted by strategy I do not match a normal probability distribution. The actual distribution obtained is (right-)skewed and indeed the mean and the median values for these two parameter distributions are not equal. Furthermore, the expectation values of the parameters as extracted by this strategy, are found to be $\bar{\varphi}_{\mathrm{i}}(T_{\mathrm{R}}) = 0.912$ and $\bar{p} = 0.461$. These values do not meet the “true”, selected, input values of these parameters, which shows that strategy I is a biased parameter extraction method. By contrast, the parameter values obtained by strategy II are normally distributed and the expectation value meets the true value of the parameters. These properties are well-explained in nonlinear regression literature [96]. Namely, under fairly general conditions, the least-squares estimates of the extracted parameters are asymptotically unbiased, have asymptotic minimum variance and are asymptotically normal. The main basis for these properties is the tangent-plane approximation for the expectation surface in the neighbourhood of the true parameter values. If this approximation is unsatisfactory, then parameter estimates will not have the three large-sample properties listed above. In other words, this property of the nonlinear

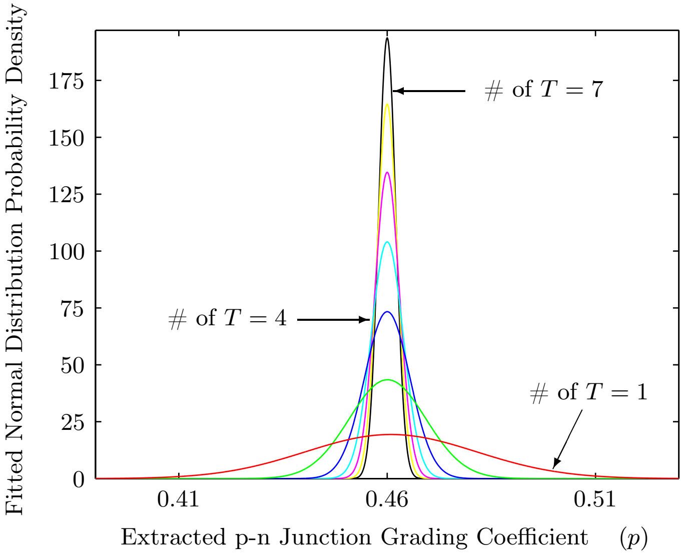

Figure 6.5 – Normal distribution probability density functions that were fitted to the histograms of estimated p-n junction grading coefficient p with different numbers of temperatures used within the extraction process. Expectably, standard deviations of statistical distributions of extracted parameter values decrease as the number of used temperatures is increased just as in figure for the estimated built-in potential.

model is visible if the noise affection is strong enough to ensure that the expectation surface can no longer be approximated by the first Taylor polynomial/series term. This implies that if the noise would be low enough the $\varphi_{\mathrm{i}}(T_{\mathrm{R}})$ and p histograms obtained by the strategy I would fit back to the normal distribution curve. Likewise if the noise variance would be increased enough, discrepancy between the histogram and for of normal probability density function would appear in strategy II as well.

Figure 6.4 and Figure 6.5 shows normal distribution fits of the built-in potential at reference temperature and grading coefficient histograms, analogous to the ones introduced in figures 6.2(b)/(c) and 6.2(e)/(f), as obtained from application of both strategy I and strategy II, the latter being applied to data sets taken on 2, 3, 4, up to 7 temperatures. As all distributions are shown on the same scale, these plots clearly show how the variance in extracted parameter values decreases with increasing number of temperatures at which data is included in the extraction effort. This is partially caused by the simple increase of the number of points. Namely, if the pure number of voltage points in strategy I is say, almost doubled (increased from 51 to 100), the extracted parameter variances would be reduced by roughly 25%. Of course this tendency saturates as the number of voltage points continues to be increased and becomes just impractical as well. However, introducing points over multiple temperatures into the data set on which parameter estimation is performed does not bring only quantitative increment but also the qualitative benefit.

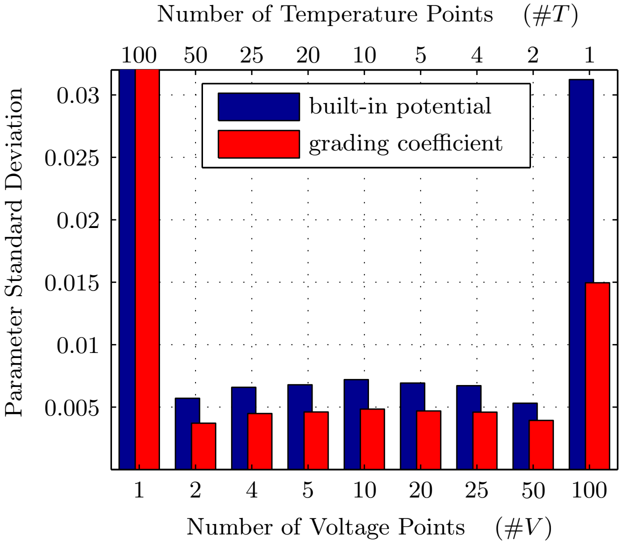

Figure 6.6 – Standard deviation of the built-in potential $\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)$ (dark bars) and standard deviation of the grading coefficient p (light bars) of 10^6 p-n junction parameters extracted from data sets generated numerically by solving equations (6.1) to (6.3) with added normally distributed pseudorandom variable to mimic stochastic measurement errors. The number of data points from which individual extraction is performed is kept constant to one hundred (100) while the ratio of voltage and temperature points was varied. Using both temperature and voltage points yielded much lower variance compared to cases in which only points of one kind are used. This demonstrates the qualitative difference between the described parameter extraction methodologies.

This claim is best illustrated through an experiment in which total number of data points is kept constant, while the proportion of voltage and temperature points in it is varied. In particular, the product of the number of voltage and the number of temperature points is fixed at one hundred. When more then one voltage point is used, the points where chosen equidistantly between −0.7 V and 0.3 V, otherwise it was chosen to be zero. Similarly, when there was more then one temperature point, temperatures were chosen in an interval between 300 K and 400 K with equidistant steps. The values of depletion capacitance model parameters were the same as in Subsection 6.4.1. Standard deviations, of extracted built-in potential at reference temperature $\varphi_{\mathrm{i}}\left(T_{\mathrm{R}}\right)$ and grading coefficient p, obtained by the each combination of number of voltage and number of temperature points which product yields a hundred, are given by the bar graph in Figure 6.6. Variant with two temperature and fifty voltage points gave the lowest deviation for the built-in potential at reference temperature, while two voltages and fifty temperatures yield the lowest grading coefficient variance. However, from Figure 6.6 it can be concluded that utilizing temperature and voltage points simultaneously is considerably favorable than putting all one hundred points to voltages or temperatures, later being extremely impractical as well.

Extraction methodology in which capacitance values at several temperatures are exploited is sensitive to the temperature scaling rule, in this particular case, on the band gap voltage. Numerical calculations have been performed and sensitivity to V_g has been confirmed. They show the tendency in which if the band gap is overestimated, all three of the junction parameters are underestimated, and vice versa. Thereby, the parameter statistical variations remains the same, just the mean values change, in other words, histograms are just shifted left or right, while their shapes remain unchanged. Anyhow, the method in which seven temperatures are used in extraction is so superior that even if all four parameters (three junction plus the band gap) are simultaneously fitted on obtained data sample, the junction parameters' variance remains lower than when extracted on only reference temperature. Even so, it could better serve as the band gap value verification possibility, rather than as a new extraction method. Since the accuracy of the estimated band gap is of such a great importance, in the next section more on the topics of its extraction can be found.

6.5 Extraction of the band gap energy/voltage

As a building block of bipolar junction and field effect transistors, but as well on its own as diode, the p-n junction is a foundation pillar of modern electronics. Therefore, the modeling of the p-n junction is of great importance for semiconductor compact models. Describing the current-voltage diode characteristic, the Shockley ideal diode equation given by (6.5), is a starting point of almost every diode compact model

$$I\left(V,T\right)=I_{\mathrm{S}}\left(T\right)\left(e^{V/V_{\mathrm{T}}(T)}-1\right)=I_{\mathrm{S}}\left(T\right)\left(e^{qV/k_{\mathrm{B}}T}-1\right) \quad . \tag{6.5}$$

The, so called, diode law, governs the dependence between current flowing through a diode I, as a function of applied voltage between anode and cathode V, where $I_{\mathrm{S}}\left(T\right)$ and $V_{\mathrm{T}}\left(T\right)$ are the temperature dependent reverse bias saturation current and the thermal voltage, respectively. The thermal voltage is linearly dependent on the absolute temperature T, through an expression $V_{\mathrm{T}}=k_{\mathrm{B}}T/q$, here k_{B} is the Boltzmann's constant and q is the elementary charge. The dependence of the ideal diode current on temperature is governed by the thermal voltage and temperature dependence of the reverse bias saturation current [101] given by

$$I_{\mathrm{S}}=I_{\mathrm{S}}\left(T_{\mathrm{R}}\right)\left(\frac{T}{T_{\mathrm{R}}}\right)^{3}e^{\frac{qV_g}{k_{\mathrm{B}}}(1/T_{\mathrm{R}}-1/T)}=I_{\mathrm{S}}\left(T_{\mathrm{R}}\right)\left(\frac{T}{T_{\mathrm{R}}}\right)^{3}e^{\frac{E_g}{k_{\mathrm{B}}}(1/T_{\mathrm{R}}-1/T)} \quad , \tag{6.6}$$

where $I_{\mathrm{S}}\left(T_{\mathrm{R}}\right)$ is value of the saturation current at reference temperature T_{R}, and E_g is the band gap whose temperature dependence is neglected in this case but can easily be included. It can be easily concluded that ideal diode characteristic over temperature is determined by the parameter $I_{\mathrm{S}}\left(T_{\mathrm{R}}\right)$ and its temperature scaling parameter V_g. The diode ideal current model given by equations (6.5) and (6.6) with slight variations is present in almost every standard semiconductor compact model. Consequently, precise extraction of the two ideal diode current parameters is essential for accurate prediction of diode ideal current as a function of voltage and temperature but as well as the depletion capacitance model which depends on the band gap energy/voltage.

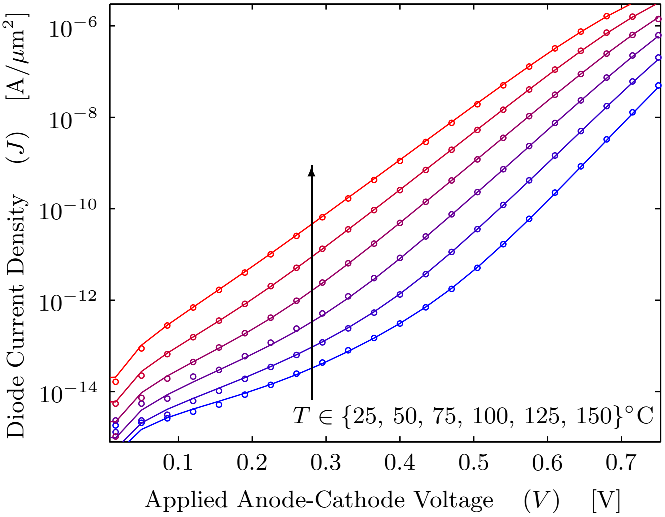

Figure 6.7 – Measured (markers) and simulated (lines) diode current I as a function of applied diode voltage V on six temperatures $T \in \{25, 50, 75, 100, 125, 150\}$°C. Measurements were done on a typical industrial p-n heterojunction that can be exploited for the base-emitter junction of the bipolar transistor. Two types of noise could be identified. One type of noise is present due to measurement equipment limits and can be observed in the part of the plot where current and voltage levels are fairly low. The other type of noise is proportional to the current value and is present all over the diode characteristics. It is identified by the nonideal fit of the simulated values.

6.5.1 Measurement noise affection

The two ideal diode current parameters are extracted utilizing nonlinear regression fitting [96] of the described model on the region of the current-voltage diode characteristics where the ideality factor is close or equal to one. Inaccuracy of the estimated model parameters is caused by noise affected measurement only. Two types of noise that are always present during diode characteristics measurement can be identified from the measured data shown in Figure 6.7. The first one, would be noise that is independent of the measured current and is present mainly due to measurement setup limits. It is usually modeled simply by adding a normally distributed random variable with zero mean and certain variance to the measured quantity. If this type of measurement noise is dominant in the part of characteristics, that part of the independent (controlled) variable can be excluded in nonlinear regression procedure. This last fact makes this noise type less important in the ideal diode current parameters extraction praxis.

The other measurement noise type is dependent on the measured variable, current density in this case. Based on experimental results, linear dispersion proportionality

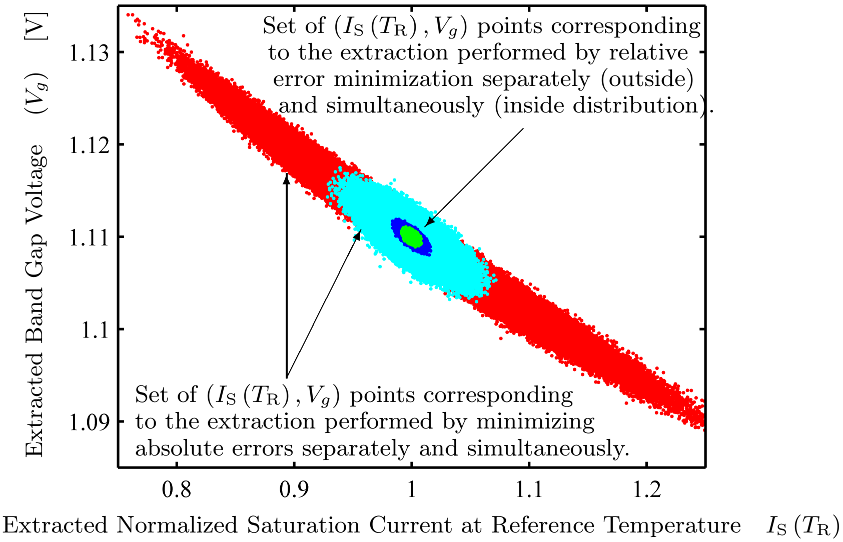

Figure 6.8 – Distribution of one million extracted pairs ($I_S(T_R)$, V_g) of the p-n junction reverse bias saturation current at reference temperature $I_S(T_R)$ and the band gap voltage V_g. Four distinct distributions correspond to extractions from the same data set only varying the employed extraction methodology. Going from highest to lowest dispersion are the absolute error minimization simultaneous and then subsequent extraction, followed by the relative error minimization separate and simultaneous parameter estimation. From the plot distributions' confidence intervals can be determined.

to the diode current density J can be assumed. As the main difference from the noise independent of diode current, effects of the noise proportional to the diode current cannot be escaped, as they are equally present all over the part of the characteristic where the nonlinear regression fitting is performed.

6.5.2 Numerical statistical experiment

Extraction of the ideal current parameters in presence of the noise proportional to the diode current can be performed by minimizing absolute or relative errors inside the least squares algorithm. Also, parameter estimation can be done by first extracting saturation current, I_S, for every available temperature and then fitting the band gap voltage, V_g, to those saturation current values, or the process of extraction of the two mentioned parameters could be performed simultaneously on current data available over voltage and temperature. The question that raises is whether these four distinct procedures yield extracted parameters of equal quality and if they do not, which one is better in terms of statistical properties like parameter's mean and variance. In order to answer the question, a numerical statistical experiment was constructed.

Description of the numerical experiment

The experiment consists of solving model equations (6.5) and (6.6) at fifty-one equidistant voltage points, in an interval between 0.25 V and 0.75 V, and at five temperatures ranging from 250 K to 450 K, where the reference temperature is chosen to be $T_{\mathrm{R}} = 300\,\mathrm{K}$, with parameters $I_{\mathrm{S}}\left(T_{\mathrm{R}}\right) = 1$ and $V_g = 1.11\,\mathrm{V}$. The reverse bias junction saturation current parameter was normalized without loss of the experiment generality. On every obtained diode current solution, normally distributed noise, $\mathcal{N}(0, 0.025)$, with zero mean and variance of 0.025 was added. Next, the parameter estimation with nonlinear regression was performed over all four methodologies. Finally, the process is repeated one million times in order to obtain a statistically significant sample.

Experiment results

Experiment results tell that extracted parameters vary very much between the four extraction procedures in terms of parameter dispersion, while at the same time mathematical expectation remains the same. Points corresponding to each extraction are shown in Figure 6.8 and can be used for confidence interval contours indication. Minimizing relative errors, expectedly, gave much better result (lower deviation) independent of whether separate or simultaneous extraction procedure is used. On the other hand the best and the worst result in terms of statistical dispersion produced relative simultaneous and absolute simultaneous extraction procedure, respectively.

6.6 Conclusion and Discussion

Two general classes of strategies, both of which are commonly applied in the industrial context, for the extraction of compact model parameters are identified in this chapter. By applying these to the example case of extraction of the p-n junction depletion capacitance, as well as the ideal diode current, compact model parameters on a typical industrial devices, it is demonstrated that the reproducibility, as quantified by the variance, of extracted parameter values can strongly depend, not only on the number of points or the point placement used within the extraction strategy, but on the extraction strategy itself. Also, from the conduced statistical experiment it may be learned that whenever there is a large change of dependent variable (for example, exponential function of the controlled variable) affected by the noise proportional to the variable, it is better to use the relative error minimization techniques. It is argued that in most of the times, lower deviation of extracted parameters are obtained when parameters are extracted simultaneously with their (temperature or geometry) scaling rule rather than separately.

The applied parameter estimation strategies are not new as such, but to my best knowledge a systematic assessment and comparison of their statistical properties and conclusions, as such have, so far not been explicitly addressed by the semiconductor technology literature. An approach to carry out such an assessment is demonstrated. This approach would be applicable quite generally, so as to enable selection of effective parameter extraction strategies for semiconductor compact models quite in general.

Chapter 7

Conclusions and Recommendations

With no obvious roadblock to further device speed enhancement, the march toward the terahertz bands with semiconductor technology will continue for the foreseeable future, opening enormous opportunities for terahertz-band applications.

In the presented thesis breakdown mechanisms have been discussed. Specifically, possible characterization techniques as well as the implications of breakdown on some important device properties and thereby the repercussions on integrated circuit design have been covered in detail. Throughout the thesis, however, most of the attention has beet paid to efficient compact modeling of breakdown phenomena for use in solid-state circuit design. First, avalanche breakdown has been introduced and its effects in multidimensional device structures are presented. This is followed by influences of impact ionization on AC bipolar transistor characteristics. Secondly, a novel model for (band-to-band) tunneling breakdown has been created for accurate simulation of junction leakage in both analog and digital circuits. Finally, parameter extraction strategies in general and estimation of p-n junction electric field parameters in particular, have been analyzed in-depth. In this chapter, specific conclusions and recommendations for potential future work related to each of these results are summarized.

7.1 Conclusions

Quasidistributed avalanche modeling in bipolar transistors

It is discussed that simple one-dimensional compact models are not able to capture true distributed nature of bipolar transistor structures especially in regimes where avalanche is pronounced. Avalanche due to lateral voltage drop can lead to the main current pinch-in in the center of the transistor. These effects can be captured by the quasidistributed transistor models with the expense paid in terms of computational cost. A simplification technique that utilizes a normalized bilinear approximation is employed for reducing model complexity. A rudiments of this method are explained in details and the model is practically implemented in Verilog-A on the basis of standard bipolar compact model. The additional parameter necessary for the full model definition is pointed and its extraction technique is portrayed. The quantitative and comparison (with currently available model) results are discussed. Model complexity is reduced from a quadratic to a linear function of the inner matrix size. The results are showing notable gains in calculation time without notable loss in the accuracy.

Avalanche breakdown of bipolar transistors in AC regime

To meet the increasing demands for high operating frequency and high output power in modern bipolar transistor applications, circuit designers explore regimes of transistor operation close to or within the avalanche breakdown region. In order to qualify and quantitatively model the effects occurring in the impact ionization regime, transistors are biased within the avalanche multiplication and small alternating signal is lead to its input. Collapse of the unilateral power gain due to impact ionization effects, as quantified by the avalanche-induced conductance is demonstrated. Physical origin of the conductance is identified and the repercussions of avalanche on the maximum available power gain, as well as on Rollett's invariant, the stability factor, is addressed. It is argued that the mentioned consequences may be explained in terms of internal negative feedback loop that is established by the impact ionization-induced conductance between output and input nodes. The frequency dependence of these quantities is described and commented in detail. The concepts and analyses are illustrated by RF measurements on modern industrial heterojunction devices and by corresponding computer simulations employing a standard compact model for bipolar transistors. Though all examples were performed measuring and simulating NPN types, all physical concepts are qualitatively and quantitatively also applicable to PNP type bipolar junction transistors. It is remarked that the effects of avalanche on AC characteristics and figures of merit may be masked by higher order effects of parasitic resistances and capacitances. However, according to the conduced analysis, trends in industry imply that described and discussed avalanche phenomena tend to be dominant over parasitic effects in the most modern and upcoming technology generations.

Compact modeling of tunneling breakdown

A novel physics-based compact model of band-to-band tunneling (BtBT) current that advances the present day state of the art models in several aspects is presented. Band-to-band tunneling may occur in highly doped, reverse biased, p-n junctions of MOS field effect, as well as in (heterojunction) bipolar transistors.

The model implements an identically vanishing BtBT current in the forward bias regime. This is in accordance with tunneling theory and attractive in terms of computational efficiency. Based on the incorporation of physical quantum states occupation effects, the model features a smooth transition, at zero bias, from the reverse bias to zero forward bias regime, thus avoiding potential simulator convergence issues.

The model includes fully physics-based temperature scaling rules that take into account temperature dependencies of both the band gap and the built-in junction electric field. In order to have the truthful modeling results, both effects indeed have to be explicitly taken into account. This statement is experimentally verified on the state of the art industrial SiGe heterojunction bipolar transistors.

The presented band-to-band tunneling model introduces a nonstraightforward parameter definition that decreases sensitivity of the extracted parameter values to stochastic errors found in measured data. This results in significantly enhanced scaling abilities of the model. A parameter extraction procedure is discussed and applied to measurements taken on the industrial SiGe devices and on the in-house Si devices.

Statistical analysis of parameter extraction procedures

Two general classes of strategies, both of which are commonly applied in the industrial context for the extraction of compact model parameters, are identified in the chapter. Applying these to the example case of extraction of the p-n junction depletion capacitance as well as the ideal diode current compact model parameters on typical industrial devices, it is demonstrated that the reproducibility, as quantified by the variance, of extracted parameter values can strongly depend not only on the number of points or the point placement used within the extraction strategy, but on the extraction strategy itself. Also, from conduced statistical experiment it may be learned that whenever there is a large change of dependent variable (exponential function of the controlled variable) affected by the noise proportional to the variable, it is better to use relative error minimization. It is argued that in most of the times, lower deviation of extracted parameters are obtained when parameters are extracted simultaneously with their scaling rule (temperature or geometry) rather than separately.

The applied parameter estimation strategies are not new as such, but to my best knowledge a systematic assessment and comparison of their statistical properties and conclusions as such, have so far not been explicitly addressed by the semiconductor technology literature. An approach to carry out such an assessment is demonstrated. This approach would be applicable quite generally, so as to enable selection of the effective parameter extraction strategies for semiconductor compact models in general.

7.2 Recommendations

Extension of avalanche multiplication current compact models

The present day compact models of avalanche current are limited to weak avalanche case. In more details, it is assumed that the charge carriers generated in a process of impact ionization do not generate extra carriers. The author has experienced cases where this condition is not satisfied and actual compact models tend to underestimate the avalanche current in certain regions of interest. This can be remedied by setting the model parameter to nonphysical values. However, this solution works on the expense of the overall fit quality. Starting from the impact ionization rates and not neglecting the secondary ionization it should be possible to derive the analytical expression for the strong avalanche model.

Quasidistributed avalanche modeling in bipolar transistors

All the phenomena described in connection to quasidistributed modeling are purely electrical. In reality, of course, thermal effects are also present. Thermal effects are especially pronounced in the avalanche regime (which is driven by relatively high electric field values) because higher current, in combination with high voltage, dissipates more power which is then converted to heat. The idea would be to to incorporate a thermal network in each of the intrinsic transistor sections and use it to study the self-heating effects combined with the distributed avalanche effects. The temperature of the bipolar device in avalanche breakdown is distributed also, since the main current flow has direct influence on local temperature. Incorporation of the thermal effects would unify three-dimensional electrical and electrothermal effects. The model would hence enable us to study the combination of both influences on device stability.

Avalanche breakdown of bipolar transistors in AC regime

The real RF power amplifiers found at transmitter output stages produce large signals at the output. This means that nonlinearities will influence model characteristics to some extent. The studies of nonlinear distortion, distribution of higher harmonics, for example are not possible in small signal regime where circuit elements are linearized. The simple AC sweep analyses should be replaced by the harmonic balance that yields steady state large signal solutions up to an arbitrary harmonic or transient ones. After an extensive simulation work, the real benefits of a large signal model verification would be achieved, of course only by verification against measurements.

In order to correctly perform large signal measurements, source and load impedances have to be known. This is truly the case only on specialized measurement setups like load-pull system, for example. Measurements should be started with single tone measurements, covering two tone ones, up to multitone measurements with real signals as found in communication band channels. After such an analysis the real potential of the models for describing large signal device performance will be visible.

Parameter extraction strategies (and their statistical analysis)

During the extraction of complete set of parameter for a given transistor, I have noticed that only the low current part is well-explained and covered in-depth. If the parameter set should fit all two-port parameters (of a small signal transistor representation) over bias *and* frequency it remains unclear how to accomplish such task or whether such task can be accomplished with a given compact model at all. I believe that deeper and more standardized procedure for extracting some nonstandard compact model parameter that eliminates this lack would be greatly appreciated by the compact modeling community. The presented statistical analysis yields useful indications about parameter extraction methodologies. Nevertheless, it might be upgraded so as to better correspond to reality. I reckon that more insight into a certain estimation strategy might be achieved by accounting for stochastic errors found in control variables (temperature and voltage in the presented cases). Besides the ones already presented, another parameter extraction procedures could be assessed.

Appendix A

Compact modeling of avalanche in bipolar transistors

Physics behind avalanche has been studied in Chapter 2. How from impact ionization rates it was possible to arrive to avalanche multiplication current was also shown. Generation coefficient 2.61 was analytically expressed utilizing several justified approximations. To repeat, the total avalanche current is the ionisation coefficient times the epilayer current, integrated over all positions where this ionisation takes place. This holds of course only in the weak avalanche regime, where the generated current does not generate extra avalanche itself. Since the avalanching is not explicitely taken into account it is a bit strange to call the whole process in compact models an avalanch, however since this terminology is widely used it is keept throughout this book. However, what is still missing in the generation coefficient are the quantities χ and $E_{\max}$ that are found in the expression. They will be calculated here.

Basic Avalanche Modeling

In this section the maximum of the electric field $E_{\max}$, and the extrapolation length χ, for the basic avalanche modeling are evaluated. In this case, the Mextram flag called "extended avalanche" has to be put to zero: EXAVL = 0. Extended avalanche modeling (EXAVL = 1) will be discussed in the next part of this appendix. As mentioned before the electric field is important for the accurate modeling of impact ionization. Schematic representation of the electric field is given in Figure A.1. The average value of the electric field over the depletion region is

$$\bar{E} = \frac{\varphi_{\mathrm{iC}} - V_{\mathrm{BC}}}{W_{\mathrm{D}}} \quad , \tag{A.1}$$

where φ_{iC} is the base-collector junction embedded voltage, W_{D} is the width of the depletion region, calculated in Appendix C. This expression implies that the average of the electric field becomes zero when $V_{\mathrm{BC}} = \varphi_{\mathrm{iC}}$. In that case the base-collector junction is already far in forward and the epilayer will be flooded with electrons and

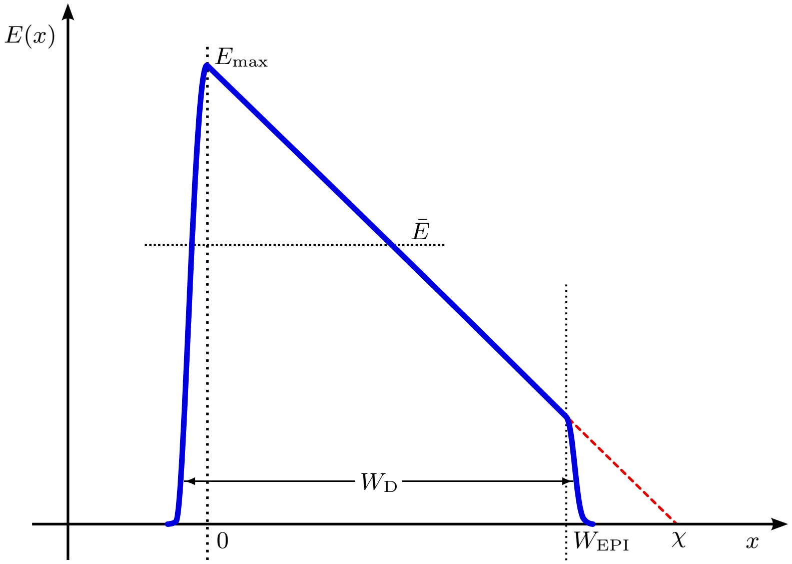

Figure A.1 – Absolute value of electric field distribution for use in the avalanche model.

holes resulting in a low electric field. Therefore, $I_{\mathrm{AVL}} = 0$ when $V_{\mathrm{BC}} \geq \varphi_{\mathrm{iC}}$. Note that the expressions below are such that also the maximum of the electric field will go to zero when its average goes to zero. The expression (2.61) for the generation factor implies that at that point also X_n will go to zero (including all its derivatives).

In the depletion region the electric field is sufficiently high to assumed that the velocity of electrons is saturated. It is assumed that in NPN transistor case there will be no holes in these regions either. The electron density, however, depends on the current density. Since the electron velocity v_{SAT} is constant $qn = |J_{\mathrm{EPI}}|/v_{\mathrm{SAT}}$, the total net charge density is then given by a sum of the doping charge and the charge density resulting from the current: $\rho = qN_{\mathrm{EPI}} - |J_{\mathrm{EPI}}|/v_{\mathrm{SAT}}$. For the charge density it does not matter whether the current moves forth or back. This gives

$$\frac{dE}{dx} = \frac{qN_{\mathrm{EPI}}}{\varepsilon_0\varepsilon_s}\left(1 - \frac{I_{\mathrm{EPI}}}{I_{\mathrm{HC}}}\right) \quad \Longrightarrow \quad \left.\frac{dE}{dx}\right|_{I_{\mathrm{EPI}}=0} = \frac{qN_{\mathrm{EPI}}}{\varepsilon_0\varepsilon_s} = \frac{2\mathrm{V}_{\mathrm{AVL}}}{W_{\mathrm{AVL}}^2} \quad , \tag{A.2}$$

where the hot-carrier current density is defined as $J_{\mathrm{HC}} = qN_{\mathrm{EPI}}v_{\mathrm{SAT}}$. If the epilayer current is equal to the hot-carrier current, the total charge in that part of the epilayer will vanish. These regions are still called depleted, since the electrons are moving with v_{SAT}, in contrast to the ohmic regions. Now the derivative of the electric field is considered. Also here the second parameter of the avalanche model $\mathrm{V}_{\mathrm{AVL}}$, is introduced. This new parameter is therefore a measure for the derivative of the electric field, especially around the maximum electric field. For this simple and one-dimensional model it should be equal to the punch-through voltage. In practice the electric field does not really have a triangular shape. Especially due to nonlocal effects the effective electric field is much broader around its maximum. This means that the value of $\mathrm{V}_{\mathrm{AVL}}$ can become small. The direct relation with the doping level is then also lost. The electric field E_0 at the base-collector junction is calculated from Figure A.1 as

$$E_0 = \bar{E} + \frac{1}{2} W_{\mathrm{D}} \left.\frac{dE}{dx}\right|_{I_{\mathrm{EPI}}=0} \left(1 - \frac{I_{\mathrm{EPI}}}{I_{\mathrm{HC}}}\right) \quad , \tag{A.3}$$

In normal operating regimes the maximum of the electric field will be at the base-collector junction, and therefore $E_{\max} = E_0$. If, due to the reversal of the slope of the electric field (Kirk effect), the maximum of the electric field moves to the epilayer-buried layer interface, the model becomes somewhat more complex and numerically more unstable. Mextram, as said, describes these effects, but will only do so when EXAVL = 1. Here only the basic model is discussed.

Next, the extrapolation length, χ, is calculated. From equation (2.59) follows

$$\left|\frac{dE}{dx}\right| = \frac{E_{\max}}{\chi} = \left.\frac{dE}{dx}\right|_{I_{\mathrm{EPI}}=0} \left(1 - \frac{I_{\mathrm{EPI}}}{I_{\mathrm{HC}}}\right) \quad . \tag{A.4}$$

Expression which can be used also in extended modeling is preferred, when the expression for the maximum electric field is modified. Electric field can be also written

$$|E\left(x\right)| = E_0 - \frac{2x}{W_{\mathrm{D}}} \left(E_0 - \bar{E}\right) \quad , \tag{A.5}$$

which is given in such a way that the electric field at the middle of depletion region $x = W_{\mathrm{D}}/2$ equals the average electric field $|E(W_{\mathrm{D}}/2)| = \bar{E}$. In the case discussed here it is always valid to make an identity $E_0 = E_{\max}$. From the expression for the electric field, and from equation $|dE/dx| = E_{\max}/\chi$, the χ is found to be

$$\chi = \frac{E_{\max} W_{\mathrm{D}}}{2\left(E_{\max} - \bar{E}\right)} \quad . \tag{A.6}$$

The same expression for χ can be found if the maximum of the electric field is at the epilayer-buried layer interface (as will be discussed next), in which case the electric field is given by (A.8).

The last what needs to be done is calculating the thickness of the depletion layer. As mentioned before, a very simple abrupt junction depletion model is used, giving

$$W_{\mathrm{D}} = \sqrt{\frac{2}{dE/dx|_{I_{\mathrm{EPI}}=0}}} \sqrt{\frac{\varphi_{\mathrm{iC}} - V_{\mathrm{BC}}}{1 - I_{\mathrm{EPI}}/I_{\mathrm{HC}}}} \quad , \tag{A.7}$$

where I_{EPI} is already defined earlier. This formula can lead to the depletion layers larger than the (effective) epilayer width W_{EFF} (here taken to be equal to W_{AVL}). Therefore the thickness over which the electric field is important is empirically shaped.

The value of X_n can not be used directly to calculate the avalanche current, because it may become very large, for instance in the iteration process of a circuit simulator, thus destroying convergency. An upper bound is considered to prevent this. There is a demand that $X_n < 1$ This means that the avalanche current can never be larger than the epilayer current, which is only a trait of the weak avalanche.

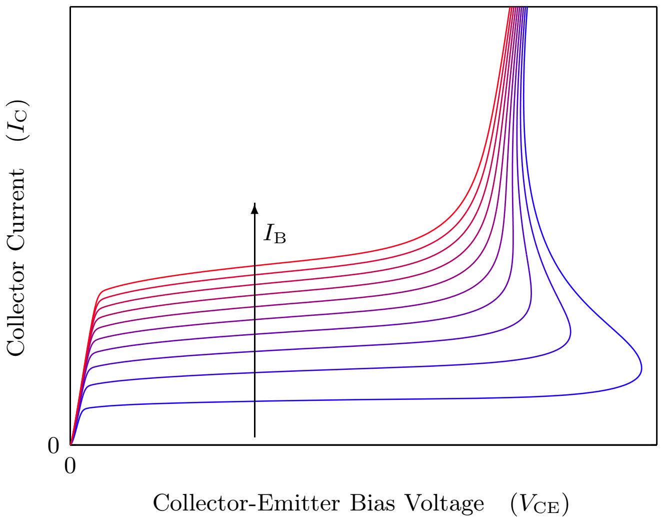

Figure A.2 – The output characteristic of a bipolar transistor showing snapback effect.

Extended Avalanche Modeling

Mextram is also equipped with an extended avalanche model. It can be switched on by setting parameter EXAVL = 1. Two extra effects are then taken into account: (i) the decrease of the effective epilayer width due to base-widening and (ii) the effect that due to a change in sign of the slope of the electric field the maximum of the electric field moves to the epilayer-buried layer interface. When these effects are included it is possible to describe snapback effects at high currents, see Figure A.2. Although this describes a physical effect, it can lead to serious convergence problems (multiple solutions are possible). It is the main reason why that part of the model is optional. Both, the effective width of the epilayer becomes smaller due to carrier injection and the electric field is different. Effective width is calculated from the equations dedicated to the epilayer current in quasisaturation, which are then numerically manipulated.

For the description of the second effect, which is of course very much related to quasisaturation and the Kirk effect, the electric field $E_W = E(W)$ at the end of the epilayer has to be calculated. Similarly to the expression for the basic avalanche modeling, the equation for the electric field for an extended avalanche modeling, from which extrapolation length in the same form is obtained, is written as

$$|E\,(x)| = E_W + 2\,(x - W_{\mathrm{D}})\left(E_W - \bar{E}\right)/W_{\mathrm{D}} \quad . \tag{A.8}$$

It is allowed for E_W to be below $\bar{E}$, whereas it is always $E_0 > \bar{E}$. For $I_{\mathrm{EPI}} = 0$ the maximum of the electric field is at the base side $E_M = E_0$.

Apart from the change in W_{EFF} and E_{max} the extended avalanche model is the same as the basic avalanche model described in the previous part of this appendix.

Appendix B

Arbitrary function smooth (C^∞) transition to a constant

From the product rule it follows that the derivative of the product of any function differentiable at a certain point and any function whose value and derivative at the same point are zero will equal zero. Therefore, all derivatives of (5.1) at zero can be made to vanish by multiplication of (5.1) by a suitable function f with property that its value and all derivatives at zero are zero. For negative arguments, say $x < 0$, the function $f(x)$ should readily approach 1. A suitable real, non-analytical yet infinitely differentiable function that has the required properties is

$$f(x) = \begin{cases} \exp\left(-\delta/x^2\right) & \text{for } x < 0 \\ 0 & \text{for } x \geq 0 \end{cases} \quad , \quad x \in \mathbb{R} \quad ; \tag{B.1}$$

the parameter $\delta > 0$ controls the speed of the transition from zero to one for $x \to -\infty$. It is straightforward to show that

$$\lim_{x\to 0^-} \frac{d^n f(x)}{dx^n} = \lim_{x\to 0^-} \sum_{i=1}^{n} c_i \delta^i \frac{\exp(-\delta/x^2)}{x^{n+2i}} = 0 \quad , \quad \forall n \in \mathbb{N} \quad , \tag{B.2}$$

where $c_i \in \mathbb{Z}$ is some integer constant. The finite power function in the denominator of the fraction under the summation approaches zero in the limit more slowly than the exponential function in numerator does. Therefore, all summands equal zero and hence the n^{th} derivative of $f(x)$ at zero vanishes. Additionally,

$$\lim_{x\to 0^-} f(x) = 0^+ \quad \text{and} \quad \lim_{x\to -\infty} f(x) = 1^- \quad . \tag{B.3}$$

Since the band-to-band tunneling current model (5.1) inherently poses exponential dependency on the applied voltage, multiplication of two exponential functions can be implemented by the addition of the exponents. The value of δ should be small enough to be neglected with respect to the rest of the argument of the exponential function

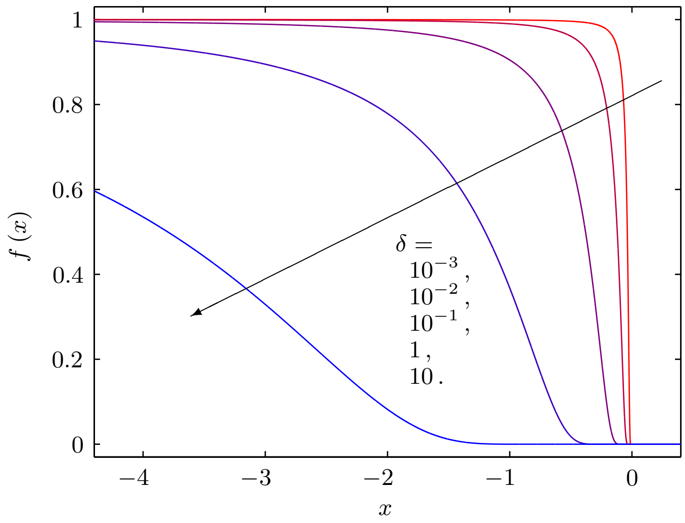

Figure B.1 – Smoothing function $f(x)$ defined by (B.1) for several values of positive parameter δ. As δ decreases transition between zero and one becomes quicker.

and it should be large enough to ensure a smooth transition within several simulation voltage sampling steps. Based on calculations it can be concluded that $\delta = 10^{-3}\,\mathrm{A}_{\mathrm{z(T)}}$ presents a reasonable compromise between the two opposing constraints.

In this appendix only a special case of an arbitrary smooth function transition to zero at zero is shown, but it is rather trivial to make it somewhat more general. Hence, shown trick may be used for implementation of smooth transition to any constant value (with certain manipulation this part can be also be made more general as to write arbitrary function) at any argument point of an arbitrary function that is of course smooth itself. Also arbitrary angle, that is the two functions can have smooth transition between each other having arbitrary first derivative value at the connection point.

Appendix C

Compact model of p-n junction electric field

In almost every semiconductor textbook there is a derivation for p-n junction's electric field, space charge region width and depletion capacitance. However, in these books only the abrupt and linear charge density profiles are addressed. In this place, the most general form of derivation is presented, taking into account the most general doping profiles from which all other derivations follow as a special cases.

Standard full depletion approximation under which is assumed that the depletion region around the metallurgical junction has well-defined edges. It also assumes that the transition between the depleted and the quasineutral region is abrupt. The quasineutral region is defined as the region adjacent to the depletion region where the electric field is small and the free carrier density is close to the net doping density. The full depletion approximation is justified by the fact that the carrier densities change exponentially with the position of the Fermi energy relative to the band edges.

Assuming one dimensional case abscissa's zero is placed at the metallurgical junction the effective doping concentrations of acceptors in the left-hand-side and donors in the right-hand side, are assumed to be

$$N(x) = \begin{cases} -N_{\mathrm{A}} \left(x/a\right)^{1/p-2} & \text{for } x < 0 \\ N_{\mathrm{D}} \left(x/d\right)^{1/p-2} & \text{for } x > 0 \end{cases} , \tag{C.1}$$

where, N_{A} and N_{A} are reference doping concentrations for acceptors and donors, a and d are the lengthscales for acceptors and donors, respectively, and p is the so called junction grading coefficient. All five are the positive real constants. Grading coefficient takes values between zero and unity and for $p = 1/2$ the consant abrupt doping profile is defined while for $p = 1/3$ the linear junction. For this case the Gauss's law in differential form reads

$$-\frac{d^2\varphi\left(x\right)}{dx^2} = \frac{dE\left(x\right)}{dx} = \frac{qN\left(x\right)}{\varepsilon_0\varepsilon_{\mathrm{s}}} \quad , \tag{C.2}$$

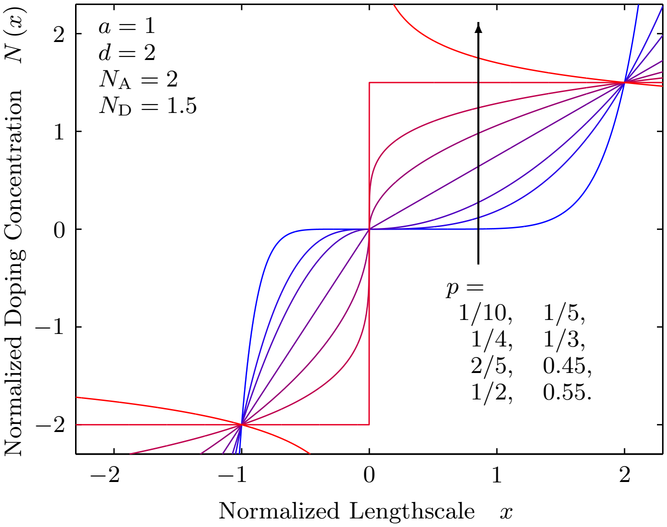

Figure C.1 – Doping concentration for several values of grading coefficient $p \in \{1/10,\ 1/5,\ 1/4,\ 1/3,\ 2/5,\ 0.45,\ 1/2,\ 0.55\}$. Metallurgical junction is at $x = 0$.

where q is the elementary charge, ε_0 is vacuum electric permittivity, ε_s is relative static permittivity of a semiconductor, E is electric field and φ is electrostatic potential. Maximum electric field value E_max is at the metallurgical junction and in monotonically decreases going toward p and n side contact. At edges between space charge region and p quasineutral region x_p and n quasineutral region x_n the electric field is zero. Maximum value of electric field equals

$$E_\mathrm{max} = E(0) = \int_{-x_p}^{0} -\frac{qN_\mathrm{A}}{\varepsilon_0\varepsilon_\mathrm{s}}\left(\frac{x}{a}\right)^{1/p-2} dx = \int_{x_n}^{0} \frac{qN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\left(\frac{x}{d}\right)^{1/p-2} dx \tag{C.3}$$

$$= \frac{qaN_\mathrm{A}}{\varepsilon_0\varepsilon_\mathrm{s}}\frac{p}{1-p}\left(\frac{-x_p}{a}\right)^{1/p-1} = -\frac{qdN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\frac{p}{1-p}\left(\frac{x_n}{d}\right)^{1/p-1} \quad . \tag{C.4}$$

From the last expression follows an identity

$$aN_\mathrm{A}\left(\frac{-x_p}{a}\right)^{1/p-1} = -dN_\mathrm{D}\left(\frac{x_n}{d}\right)^{1/p-1} \quad , \tag{C.5}$$

from which subsequently follows the ratio between the two depletion layer widths

$$\frac{x_p}{x_n} = \left(\frac{d}{a}\right)^{\frac{2p-1}{1-p}}\left(\frac{N_\mathrm{D}}{N_\mathrm{A}}\right)^{\frac{p}{1-p}} \quad , \tag{C.6}$$

about the connection between the right and left ends of space charge region.

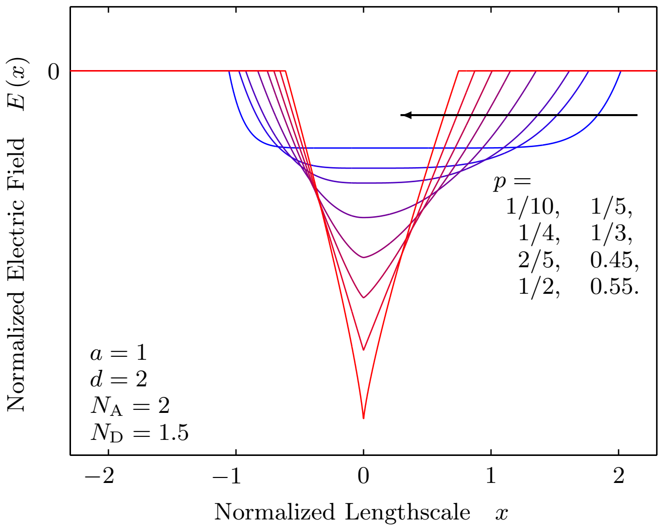

Figure C.2 – Equilibrium electric field distribution in a p-n junction for several different values of the junction grading coefficient p. Metallurgical junction is located at $x = 0$. Built-in potential φ_i is assumed to be equal for all grading coefficient values, therefrom surfaces defined by separate curves in the plot are also equal between each other.

One-dimensional Poisson's equation in an integral form of a p-n junction reads

$$-\int\limits_{-x_p}^{x_n} E(x)\;dx = -\int\limits_{-x_p}^{0} E(x)\;dx - \int\limits_{0}^{x_n} E(x)\;dx = \varphi_\mathrm{i} - V \quad , \tag{C.7}$$

where φ_i is the p-n junction built-in potential and V is the applied voltage positive for forward bias. The last expression is easily decomposed

$$\int\limits_{-x_p}^{0}\int\limits_{-x_p}^{x} \frac{qN_\mathrm{A}}{\varepsilon_0\varepsilon_\mathrm{s}}\left(\frac{x}{a}\right)^{1/p-2} dx\,dx - \int\limits_{0}^{x_n}\int\limits_{x_n}^{x} \frac{qN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\left(\frac{x}{d}\right)^{1/p-2} dx\,dx = \varphi_\mathrm{i} - V \tag{C.8}$$

and then solved yielding

$$a^{2-1/p}N_\mathrm{A}\left(-x_p\right)^{1/p} + d^{2-1/p}N_\mathrm{D}x_n^{1/p} = \frac{\varepsilon_0\varepsilon_\mathrm{s}}{qp}\left(\varphi_\mathrm{i} - V\right) \quad . \tag{C.9}$$

Equations (C.6) and (C.9) make a system of two equations. Solving it by elimination of one variable, say x_n the following expression is achieved

$$\left[1 + \left(\frac{a}{d}\right)^{\frac{2p-1}{1-p}}\left(\frac{N_\mathrm{A}}{N_\mathrm{D}}\right)^{\frac{p}{1-p}}\right]^p x_p = a^{1-2p}\left[\frac{\varepsilon_0\varepsilon_\mathrm{s}}{qpN_\mathrm{A}}\left(\varphi_\mathrm{i} - V\right)\right]^p \quad . \tag{C.10}$$

Analogously, an expression for left border of the depletion region would write

$$\left[1+\left(\frac{d}{a}\right)^{\frac{2p-1}{1-p}}\left(\frac{N_{\mathrm{D}}}{N_{\mathrm{A}}}\right)^{\frac{p}{1-p}}\right]^{p} x_n = d^{1-2p}\left[\frac{\varepsilon_0\varepsilon_{\mathrm{s}}}{qpN_{\mathrm{D}}}\left(\varphi_{\mathrm{i}}-V\right)\right]^{p} \quad . \tag{C.11}$$

Taking $p = 1/2$ for check standard p-n junction textbook formulae emerge

$$x_p = \sqrt{\frac{2\varepsilon_0\varepsilon_{\mathrm{s}}}{q}\frac{N_{\mathrm{D}}}{N_{\mathrm{A}}}\frac{\varphi_{\mathrm{i}}-V}{N_{\mathrm{D}}+N_{\mathrm{A}}}} \quad \text{and} \quad x_n = \sqrt{\frac{2\varepsilon_0\varepsilon_{\mathrm{s}}}{q}\frac{N_{\mathrm{A}}}{N_{\mathrm{D}}}\frac{\varphi_{\mathrm{i}}-V}{N_{\mathrm{A}}+N_{\mathrm{D}}}} \quad . \tag{C.12}$$

Total depletion layer width is the sum of left and right depletion edges

$$x_d = x_p + x_n = \left(\frac{\varepsilon_0\varepsilon_{\mathrm{s}}}{qp}\varphi_{\mathrm{i}}\right)^{p}\left(a^{\frac{2p-1}{p-1}}N_{\mathrm{A}}^{\frac{p}{p-1}}+d^{\frac{2p-1}{p-1}}N_{\mathrm{D}}^{\frac{p}{p-1}}\right)^{1-p}\left(1-\frac{V}{\varphi_{\mathrm{i}}}\right)^{p} \quad , \tag{C.13}$$

from which the total space charge region width as a function of voltage is

$$x_d\left(V\right) = x_d\left(0\right)\left(1-V/\varphi_{\mathrm{i}}\right)^{p} \quad . \tag{C.14}$$

Now substituting expressions (C.10) and (C.11) in (C.4) the full expression for the maximum value of junction's electric field is

$$E_{\max} = \frac{\varphi_{\mathrm{i}}^{1-p}}{1-p}\left(\frac{qp}{\varepsilon_0\varepsilon_{\mathrm{s}}}\right)^{p}\left(a^{\frac{2p-1}{p-1}}N_{\mathrm{A}}^{\frac{p}{p-1}}+d^{\frac{2p-1}{p-1}}N_{\mathrm{D}}^{\frac{p}{p-1}}\right)^{p-1}\left(1-\frac{V}{\varphi_{\mathrm{i}}}\right)^{1-p} \quad . \tag{C.15}$$

Thereby maximum electric field can be written in a more simple form

$$E_{\max}\left(V\right) = E_{\max}\left(0\right)\left(1-V/\varphi_{\mathrm{i}}\right)^{1-p} \quad , \tag{C.16}$$

The total electric charge surface density Q' within the depletion region would be

$$Q'_{\mathrm{t}} = q\int\limits_{-x_p}^{x_n}\left|N\left(x\right)\right|dx = q\int\limits_{-x_p}^{0}\left|-N_{\mathrm{A}}\left(\frac{x}{a}\right)^{1/p-2}\right|dx + q\int\limits_{0}^{x_n}N_{\mathrm{D}}\left(\frac{x}{d}\right)^{1/p-2}dx \quad , \tag{C.17}$$

where the absolute value sign is added in the definition so that either the positive or the negative charge can be used in the calculation, as they are equal in magnitude. Solving the above integrals and again substituting expressions (C.10) and (C.11) in a few lines the result writes

$$Q'_{\mathrm{t}} = 2\varepsilon_0\varepsilon_{\mathrm{s}}\frac{\varphi_{\mathrm{i}}^{1-p}}{1-p}\left(\frac{qp}{\varepsilon_0\varepsilon_{\mathrm{s}}}\right)^{p}\left(a^{\frac{2p-1}{p-1}}N_{\mathrm{A}}^{\frac{p}{p-1}}+d^{\frac{2p-1}{p-1}}N_{\mathrm{D}}^{\frac{p}{p-1}}\right)^{p-1}\left(1-\frac{V}{\varphi_{\mathrm{i}}}\right)^{1-p} \quad . \tag{C.18}$$

Comparing (C.15) with (C.18) it can be written

$$Q'_{\mathrm{t}} = Q'_{\mathrm{t}}\left(0\right)\left(1-V/\varphi_{\mathrm{i}}\right)^{1-p} = 2\varepsilon_0\varepsilon_{\mathrm{s}}E_{\max}\left(0\right)\left(1-V/\varphi_{\mathrm{i}}\right)^{1-p} \quad . \tag{C.19}$$

Net positive or negative (equal in magnitude) charge surface density change Q' when voltage V is moved from zero can be written as

$$Q' = Q'_{\mathrm{t}}(0)/2 - Q'_{\mathrm{t}}(0)\left(1 - V/\varphi_{\mathrm{i}}\right)^{1-p}/2 = Q'_{\mathrm{t}}(0)\left[1 - \left(1 - V/\varphi_{\mathrm{i}}\right)^{1-p}\right] \quad , \qquad \text{(C.20)}$$

where $Q'(0) \equiv Q'_{\mathrm{t}}(0)/2 = \varepsilon_0\varepsilon_{\mathrm{s}}E_{\max}(0)$. In order to obtain depletion layer surface capacitance density we have to find derivative with respect to applied voltage

$$C'_{\mathrm{j}}(V) \triangleq \left|\frac{dQ'(V)}{dV}\right| = \frac{(1-p)\,Q'(0)}{\varphi_{\mathrm{i}}}\left(1 - V/\varphi_{\mathrm{i}}\right)^{-p} = \frac{C'_{\mathrm{j}}(0)}{\left(1 - V/\varphi_{\mathrm{i}}\right)^{p}} \quad , \qquad \text{(C.21)}$$

where the absolute value sign is used from the same reason as in previous place. Zero bias depletion capacitance can be expressed in terms of other quantities

$$C'_{\mathrm{j}}(0) = \varepsilon_0\varepsilon_{\mathrm{s}}\left(\frac{qp}{\varepsilon_0\varepsilon_{\mathrm{s}}\varphi_{\mathrm{i}}}\right)^{p}\left(a^{\frac{2p-1}{p-1}}N_{\mathrm{A}}^{\frac{p}{p-1}} + d^{\frac{2p-1}{p-1}}N_{\mathrm{D}}^{\frac{p}{p-1}}\right)^{p-1} \quad . \qquad \text{(C.22)}$$

Note that for the sake of correctness check and completeness

$$C'_{\mathrm{j}}(0) = \varepsilon_0\varepsilon_{\mathrm{s}}/x_d(0) \quad \Longrightarrow \quad C'_{\mathrm{j}}(V) = \varepsilon_0\varepsilon_{\mathrm{s}}/x_d(V) \quad , \qquad \text{(C.23)}$$

which is just an expression for a capacitance surface density of a parallel plate capacitor, the result which could be expected. Other zero bias quantities are expressed in terms of zero bias depletion capacitance as

$$x_d(0) = \frac{\varepsilon_0\varepsilon_{\mathrm{s}}}{C'_{\mathrm{j}}(0)} \quad , \quad E_{\max}(0) = \frac{\varphi_{\mathrm{i}}C'_{\mathrm{j}}(0)}{\varepsilon_0\varepsilon_{\mathrm{s}}(1-p)} \quad , \quad Q'(0) = \frac{\varphi_{\mathrm{i}}C'_{\mathrm{j}}(0)}{1-p} \quad , \qquad \text{(C.24)}$$

which makes the compact model of the p-n junction electric field complete and self-consistent with the compact model of depletion capacitance.

What has been derived would be the most general case that could be derived analytically. Special cases of a symmetrical junction or one-sided Schottky junction are obtained in the respective parameter definitions and limits.

References

[1] J. C. Maxwell, "A dynamical theory of the electromagnetic field," *Philosophical Transactions of the Royal Society of London*, vol. 155, pp. 459–512, 1865.

[2] N. Tesla, "System of transmission of electrical energy," Patent 645 576, Mar. 20, 1900.

[3] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[4] P. H. Siegel, "Terahertz technology," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 50, no. 3, pp. 910–928, Mar. 2002.

[5] S. Lee, B. Jagannathan, S. Narasimha, A. Chou, N. Zamdmer, J. Johnson, R. Williams, L. Wagner, J. Kim, J.-O. Plouchart, J. Pekarik, S. Springer, and G. Freeman, "Record RF performance of 45-nm SOI CMOS technology," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, Dec. 2007, pp. 255–258.

[6] J.-S. Rieh, B. Jagannathan, H. Chen, K. Schonenberg, D. Angell, A. Chinthakindi, J. Florkey, F. Golan, D. Greenberg, S.-J. Jeng, M. Khater, F. Pagette, C. Schnabel, P. Smith, A. Stricker, K. Vaed, R. Volant, D. Ahlgren, G. Freeman, K. Stein, and S. Subbanna, "SiGe HBTs with cut-off frequency of 350 GHz," in *Electron Devices Meeting, 2002. IEDM '02. Digest. International*, 2002, pp. 771–774.

[7] B. Heinemann, R. Barth, D. Bolze, J. Drews, P. Formanek, T. Grabolla, U. Haak, W. Hoppner, D. Kopke, B. Kuck, R. Kurps, S. Marschmeyer, H. Richter, H. Rucker, P. Schley, D. Schmidt, W. Winkler, D. Wolansky, H. Wulf, and Y. Yamamoto, "A low-parasitic collector construction for high-speed SiGe:C HBTs," in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, Dec. 2004, pp. 251–254.

[8] J.-S. Rieh, D. Greenberg, M. Khater, K. Schonenberg, S.-J. Jeng, F. Pagette, T. Adam, A. Chinthakindi, J. Florkey, B. Jagannathan, J. Johnson, R. Krishnasamy, D. Sanderson, C. Schnabel, P. Smith, A. Stricker, S. Sweeney, K. Vaed, T. Yanagisawa, D. Ahlgren, K. Stein, and G. Freeman, "SiGe HBTs for millimeter-wave applications with simultaneously optimized fT and fmax of 300 GHz," in *Radio Frequency Integrated Circuits (RFIC) Symposium, 2004. Digest of Papers. 2004 IEEE*, Jun. 2004, pp. 395–398.

[9] R. M. Warner Jr. and R. D. Schrimpf, "Bjt-mosfet transconductance comparisons," *Electron Devices, IEEE Transactions on*, vol. 34, no. 5, pp. 1061–1065, May 1987.

[10] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 5th ed. New York: John Wiley & Sons, 2009.

[11] A. J. Joseph, D. L. Harame, B. Jagannathan, D. Coolbaugh, D. Ahlgren, J. Magerlein, L. Lanzerotti, N. Feilchenfeld, S. St Onge, J. Dunn, and E. Nowak, "Status and direction of communication technologies - SiGe BiCMOS and RFCMOS," *Proceedings of the IEEE*, vol. 93, no. 9, pp. 1539–1558, Sep. 2005.

[12] P. A. H. Hart, *Bipolar and bipolar-MOS integration.* Amsterdam: Elsevier, 1994.

[13] J. Dunn, D. L. Harame, A. J. Joseph, S. A. St. Onge, N. B. Feilchenfeld, L. Lanzerotti, B. Orner, E. Gebreselasie, J. B. Johnson, D. D. Coolbaugh, R. Rassel, and M. Khater, "SiGe BiCMOS trends - today and tomorrow," in *Custom Integrated Circuits Conference, CICC '06. IEEE*, Sep. 2006, pp. 695–702.

[14] J. M. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits*, 3rd ed. New Jersy: Prentice Hall, 2009.

[15] S. P. Voinigescu, T. O. Dickson, R. Beerkens, I. Khalid, and P. Westergaard, "A comparison of Si CMOS, SiGe BiCMOS, and InP HBT technologies for high-speed and millimeter-wave ICs," in *Silicon Monolithic Integrated Circuits in RF Systems, 2004. Digest of Papers. 2004 Topical Meeting on*, Sep. 2004, pp. 111–114.

[16] J. D. Cressler, *Silicon heterostructure handbook: materials, fabrication, devices, circuits, and applications of SiGe and Si strained-layer epitaxy*, J. D. Cressler, Ed. Boca Raton, FL: CRC Press, 2006.

[17] K. Nellis and P. J. Zampardi, "A comparison of linear handset power amplifiers in different bipolar technologies," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 10, pp. 1746–1754, Oct. 2004.

[18] J. D. Cressler and G. Niu, *Silicon-germanium heterojunction bipolar transistors.* Norwood, MA: Artech House, 2003.

[19] M. Racanelli and P. Kempf, "SiGe BiCMOS technology for RF circuit applications," *Electron Devices, IEEE Transactions on*, vol. 52, no. 7, pp. 1259–1270, Jul. 2005.

[20] H. Wong, "Drain breakdown in submicron MOSFETs: a review," *Microelectronics Reliability*, vol. 40, no. 1, pp. 3–15, Jan. 2000.

[21] A. Acovic, G. L. Rosa, and Y.-C. Sun, "A review of hot-carrier degradation mechanisms in MOSFETs," *Microelectronics and Reliability*, vol. 36, no. 7-8, pp. 845–869, 1996.

[22] H. Wong, Y. Fu, J. Liou, and Y. Yue, "Hot-carrier reliability and breakdown characteristics of multi-finger RF MOS transistors," *Microelectronics Reliability*, vol. 49, no. 1, pp. 13–16, Jan. 2009.

[23] L. E. Larson, "Silicon technology tradeoffs for radio-frequency/mixed-signal "systems-on-a-chip"," *Electron Devices, IEEE Transactions on*, vol. 50, no. 3, pp. 683–699, Mar. 2003.

[24] E. O. Johnson, "Physical limitations on frequency and power parameters of transistors," *RCA Review*, vol. 26, pp. 163–177, Jun. 1965.

[25] K. K. Ng, M. R. Frei, and C. A. King, "Reevaluation of the $f_{\mathrm{T}} \cdot \mathrm{BV}_{\mathrm{CEO}}$ limit on Si bipolar transistors," *Electron Devices, IEEE Transactions on*, vol. 45, no. 8, pp. 1854–1855, Aug. 1998.

[26] D. L. Harame, D. C. Ahlgren, D. D. Coolbaugh, J. S. Dunn, G. G. Freeman, J. D. Gillis, R. A. Groves, G. N. Hendersen, R. A. Johnson, A. J. Joseph, S. Subbanna, A. M. Victor, K. M. Watson, C. S. Webster, and P. J. Zampardi, "Current status and future trends of SiGe BiCMOS technology," *Electron Devices, IEEE Transactions on*, vol. 48, no. 11, pp. 2575–2594, Nov. 2001.

[27] J. B. Johnson, A. J. Joseph, D. C. Sheridan, R. M. Maladi, P.-O. Brandt, J. Persson, J. Andersson, A. Bjorneklett, U. Persson, F. Abasi, and L. Tilly, "Silicon-germanium BiCMOS HBT technology for wireless power amplifier applications," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 10, pp. 1605–1614, Oct. 2004.

[28] A. Inoue, S. Nakatsuka, R. Hattori, and Y. Matsuda, "The maximum operating region in sige hbts for rf power amplifiers," in *Microwave Symposium Digest, 2002 IEEE MTT-S International*, 2002, pp. 1023–1026.

[29] A. Huang and B. Zhang, "The future of bipolar power transistors," *Electron Devices, IEEE Transactions on*, vol. 48, no. 11, pp. 2535–2543, Nov. 2001.

[30] J. D. Hayden, D. Burnett, and J. Nangle, "A comparison of base current reversal and bipolar snapback in advanced n-p-n bipolar transistors," *Electron Device Letters, IEEE*, vol. 12, no. 8, pp. 407–409, Aug. 1991.

[31] M. Rickelt, H.-M. Rein, and E. Rose, "Influence of impact-ionization-induced instabilities on the maximum usable output voltage of Si-bipolar transistors," *Electron Devices, IEEE Transactions on*, vol. 48, no. 4, pp. 774–783, Apr. 2001.

[32] S. Marjanović, *Elektronika linearnih kola i sistema.* Beograd: Akademska misao, 2002.

[33] J.-S. Rieh, B. Jagannathan, D. R. Greenberg, M. Meghelli, A. Rylyakov, F. Guarin, Z. Yang, D. C. Ahlgren, G. Freeman, P. Cottrell, and D. Harame, "SiGe heterojunction bipolar transistors and circuits toward terahertz communication applications," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 52, no. 10, pp. 2390–2408, Oct. 2004.

[34] M. Levinshtein, J. Kostamovaara, and S. Vainshtein, *Breakdown phenomena in semiconductors and semiconductor devices.* Singapore: World Scientific, 2005.

[35] *International Technology Roadmap for Semiconductors*, 2009, executive summary. [Online]. Available: http://www.itrs.net

[36] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics–i. MOS technology," *Solid-State Electronics*, vol. 15, no. 7, pp. 819–829, Jul. 1972.

[37] A. Neugroschel, C.-T. Sah, and M. S. Carroll, "Degradation of bipolar transistor current gain by hot holes during reverse emitter-base bias stress," *Electron Devices, IEEE Transactions on*, vol. 43, no. 8, pp. 1286–1290, Aug. 1996.

[38] M. Reisch, *High-frequency bipolar transistors: physics, modelling, applications.* Berlin: Springer, 2003.

[39] C. C. McAndrew, "Practical modeling for circuit simulation," *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 3, pp. 439–448, Mar. 1998.

[40] H. C. de Graaff and F. M. Klaassen, *Compact transistor modelling for circuit design.* Berlin: Springer Verlag, 1990.

[41] N. Arora, *MOSFET modeling for VLSI simulation: theory and practice.* Singapore: World Scientific, 2007.

[42] R. C. Dorf, *The Electrical Engineering Handbook*, 2nd ed., R. C. Dorf, Ed. Boca Raton, FL: CRC Press, 1997.

[43] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly Journal of Applied Mathmatics*, vol. 2, no. 2, pp. 164–168, 1944.

[44] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[45] J. R. Long, "SiGe radio frequency ICs for low-power portable communication," *Proceedings of the IEEE*, vol. 93, no. 9, pp. 1598–1623, Sep. 2005.

[46] H. Gummel and H. Poon, "A compact bipolar transistor model," in *Solid-State Circuits Conference. Digest of Technical Papers. 1970 IEEE International*, vol. XIII, Feb. 1970, pp. 78–79.

[47] C. McAndrew, J. Seitchik, D. Bowers, M. Dunn, M. Foisy, I. Getreu, M. McSwain, S. Moinian, J. Parker, P. van Wijnen, and L. Wagner, "VBIC95: An improved vertical, IC bipolar transistor model," in *Bipolar/BiCMOS Circuits and Technology Meeting, 1995., Proceedings of the 1995*, Oct. 1995, pp. 170–177.

[48] C. C. McAndrew, J. A. Seitchik, D. F. Bowers, M. Dunn, M. Foisy, I. Getreu, M. McSwain, S. Moinian, J. Parker, D. J. Roulston, M. Schroter, P. van Wijnen, and L. F. Wagner, "VBIC95, the vertical bipolar inter-company model," *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 10, pp. 1476–1483, Oct. 1996.

[49] M. Schröter, "Staying current with HICUM," *Circuits and Devices Magazine, IEEE*, vol. 18, no. 3, pp. 16–25, May 2002.

[50] Mextram (in Verilog-A) homepage. [Online]. Available: http://mextram.ewi.tudelft.nl/

[51] Compact Model Council (CMC) old homepage. [Online]. Available: http://www.eigroup.org/cmc/

[52] Compact Model Council (CMC) homepage. [Online]. Available: www.geia.org/index.asp?bid=597

[53] W. J. Kloosterman and H. C. de Graaff, "Avalanche multiplication in a compact bipolar transistor model for circuit simulation," in *Bipolar Circuits and Technology Meeting, 1988., Proceedings of the 1988*, Sep. 1988, pp. 103–106.

[54] W. J. Kloosterman and H. C. De Graaff, "Avalanche multiplication in a compact bipolar transistor model for circuit simulation," *Electron Devices, IEEE Transactions on*, vol. 36, no. 7, pp. 1376–1380, Jul. 1989.

[55] M. Schröter, Z. Yan, T. Y. Lee, and W. Shi, "A compact tunneling current and collector breakdown model," in *Bipolar/BiCMOS Circuits and Technology Meeting, 1998. Proceedings of the 1998*, Sep. 1998, pp. 203–206.

[56] W. J. Kloosterman, J. C. J. Paasschens, and R. J. Havens, "A comprehensive bipolar avalanche multiplication compact model for circuit simulation," in *Bipolar/BiCMOS Circuits and Technology Meeting, 2000. Proceedings of the 2000*, 2000, pp. 172–175.

[57] J. W. Slotboom, G. Streutker, M. J. van Dort, P. H. Woerlee, A. Pruijmboom, and D. J. Gravesteijn, "Non-local impact ionization in silicon devices," in *Electron Devices Meeting, 1991. IEDM '91. Technical Digest., International*, Dec. 1991, pp. 127–130.

[58] G.-B. Hong and J. G. Fossum, "Implementation of nonlocal model for impact-ionization current in bipolar circuit simulation and application to SiGe HBT design optimization," *Electron Devices, IEEE Transactions on*, vol. 42, no. 6, pp. 1166–1173, Jun. 1995.

[59] M. Rickelt and H.-M. Rein, "Impact-ionization induced instabilities in high-speed bipolar transistors and their influence on the maximum usable output voltage," in *Bipolar/BiCMOS Circuits and Technology Meeting, 1999. Proceedings of the 1999*, 1999, pp. 54–57.

[60] ——, "A novel transistor model for simulating avalanche-breakdown effects in Si bipolar circuits," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 9, pp. 1184–1197, Sep. 2002.

[61] R. van der Toorn, J. Dohmen, and O. Hubert, "Distribution of the collector resistance of planar bipolar transistors: Impact on small signal characteristics and compact modeling," in *Bipolar/BiCMOS Circuits and Technology Meeting, 2007. BCTM '07. IEEE*, Oct. 2007, pp. 184–187.

[62] M. P. van der Heijden, H. C. de Graaff, and L. C. N. de Vreede, "A novel frequency-independent third-order intermodulation distortion cancellation technique for BJT amplifiers," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 9, pp. 1176–1183, Sep. 2002.

[63] M. Spirito, L. C. N. de Vreede, L. K. Nanver, S. Weber, and J. N. Burghartz, "Power amplifier PAE and ruggedness optimization by second-harmonic control," *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 9, pp. 1575–1583, Sep. 2003.

[64] M. Marchetti, M. J. Pelk, K. Buisman, W. Neo, M. Spirito, and L. de Vreede, "Active harmonic load-pull with realistic wideband communications signals," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 56, no. 12, pp. 2979–2988, Dec. 2008.

[65] S. G. Narendra and A. P. Chandrakasan, *Leakage in nanometer CMOS technologies*, S. G. Narendra and A. P. Chandrakasan, Eds. Basel: Birkhauser Verlag, 2006.

[66] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau, "Transistor elements for 30nm physical gate lengths and beyond," *Intel Technology Journal*, vol. 6, no. 2, pp. 42–54, May 2002.

[67] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *Electron Devices Meeting, 1998. IEDM '98 Technical Digest., International*, Dec. 1998, pp. 789 –792.

[68] H. Ananthan, A. Bansal, and K. Roy, "Analysis of drain-to-body band-to-band tunneling in double gate MOSFET," in *SOI Conference, 2005. Proceedings. 2005 IEEE International*, Oct. 2005, pp. 159–160.

[69] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, "A review of hydrodynamic and energy-transport models for semiconductor device simulation," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 251–274, Feb. 2003.

[70] R. V. Overstraeten and H. D. Man, "Measurement of the ionization rates in diffused silicon p-n junctions," *Solid-State Electronics*, vol. 13, no. 5, pp. 583–608, 1970.

[71] A. G. Chynoweth, "Ionization rates for electrons and holes in silicon," *Physical Review*, vol. 109, no. 5, pp. 1537–1540, Mar. 1958.

[72] E. Schöll, *Nonlinear spatio-temporal dynamics and chaos in semiconductors.* Cambridge: Cambridge University Press, 2001.

[73] A. Komijani and A. Hajimiri, "A wideband 77-GHz, 17.5-dBm fully integrated power amplifier in silicon," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 8, pp. 1749–1756, Aug. 2006.

[74] V. Jain, F. Tzeng, L. Zhou, and P. Heydari, "A single-chip dual-band 22-29-GHz/77-81-GHz BiCMOS transceiver for automotive radars," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 12, pp. 3469–3485, Dec. 2009.

[75] G. Freeman, M. Meghelli, Y. Kwark, S. Zier, A. Rylyakov, M. A. Sorna, T. Tanji, O. M. Schreiber, K. Walter, J.-S. Rieh, B. Jagannathan, A. Joseph, and S. Subbanna, "40-Gb/s circuits built from a 120-GHz f_T SiGe technology," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 9, pp. 1106–1114, Sep. 2002.

[76] H. Veenstra, G. A. M. Hurkx, D. van Goor, H. Brekelmans, and J. R. Long, "Analyses and design of bias circuits tolerating output voltages above $\mathrm{BV_{CEO}}$," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 10, pp. 2008–2018, Oct. 2005.

[77] P. Deixler, A. Rodriguez, W. De Boer, H. Sun, R. Colclaser, D. Bower, N. Bell, A. Yao, R. Brock, Y. Bouttement, G. A. M. Hurkx, L. F. Tiemeijer, J. C. J. Paasschens, H. G. A. Huizing, D. M. H. Hartskeerl, P. Agrarwal, P. H. C. Magnee, E. Aksen, and J. W. Slotboom, "QUBiC4X: An $f_T/f_{\max} = 130/140\,\mathrm{GHz}$ SiGe:C-BiCMOS manufacturing technology with elite passives for emerging microwave applications," in *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting, IEEE BCTM'04*, Sep. 2004, pp. 233–236.

[78] S. Mason, "Pover gain in feedback amplifier," *Circuit Theory, Transactions of the IRE Professional Group on*, vol. 1, no. 2, pp. 20–25, Jun. 1954.

[79] M. S. Gupta, "Power gain in feedback amplifiers, a classic revisited," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 40, no. 5, pp. 864–879, May 1992.

[80] J. Rollett, "Stability and power-gain invariants of linear twoports," *Circuit Theory, IRE Transactions on*, vol. 9, no. 1, pp. 29–32, Mar. 1962.

[81] P. M. Solomon, D. J. Frank, J. Jopling, C. D'Emic, O. Dokumaci, P. Ronsheim, and W. E. Haensch, "Tunnel current measurements on P/N junction diodes and implications for future device design," in *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International*, Dec. 2003, pp. 931–934.

[82] P. M. Solomon, J. Jopling, D. J. Frank, C. D'Emic, O. Dokumaci, P. Ronsheim, and W. E. Haensch, "Universal tunneling behavior in technologically relevant P/N junction diodes," *Journal of Applied Physics*, vol. 95, no. 10, pp. 5800–5812, 2004.

[83] J. D. Cressler, "On the potential of SiGe HBTs for extreme environment electronics," *Proceedings of the IEEE*, vol. 93, no. 9, pp. 1559–1582, Sep. 2005.

[84] J. Yuan, J. D. Cressler, R. Krithivasan, T. Thrivikraman, M. H. Khater, D. C. Ahlgren, A. J. Joseph, and J.-S. Rieh, "On the performance limits of cryogenically operated SiGe HBTs and its relation to scaling for terahertz speeds," *Electron Devices, IEEE Transactions on*, vol. 56, no. 5, pp. 1007–1019, May 2009.

[85] A. J. Scholten, G. D. J. Smit, M. Durand, R. van Langevelde, and D. B. M. Klaassen, "The physical background of JUNCAP2," *Electron Devices, IEEE Transactions on*, vol. 53, no. 9, pp. 2098–2107, Sep. 2006.

[86] G. A. M. Hurkx, "On the modelling of tunnelling currents in reverse-biased p-n junctions," *Solid-State Electronics*, vol. 32, no. 8, pp. 665–668, 1989.

[87] E. Kane, "Zener tunneling in semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 12, no. 2, pp. 181–188, 1960.

[88] J. L. Moll, *Physics of semiconductors.* Boston, MA: McGraw-Hill, 1964.

[89] S. Wang, *Solid-state electronics.* Boston, MA: McGraw-Hill, 1966.

[90] K. K. N. Simon M. Sze, *Physics of semiconductor devices*, 3rd ed. New York: John Wiley & Sons, 2007.

[91] E. O. Kane, "Theory of tunneling," *Journal of Applied Physics*, vol. 32, no. 1, pp. 83–91, 1961.

[92] V. Alex, S. Finkbeiner, and J. Weber, "Temperature dependence of the indirect energy gap in crystalline silicon," *Journal of Applied Physics*, vol. 79, no. 9, pp. 6943–6946, 1996.

[93] W. Yang *et al.* BSIM4.6.5 MOSFET model user's manual. [Online]. Available: http://www-device.eecs.berkeley.edu/

[94] H. J. Mattausch, M. Miura-Mattausch, N. Sadachika, M. Miyake, and D. Navarro, "The HiSIM compact model family for integrated devices containing a surface-potential MOSFET core," in *Mixed Design of Integrated Circuits and Systems, 2008. MIXDES 2008. 15th International Conference on*, Jun. 2008, pp. 39–50.

[95] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *Electron Devices, IEEE Transactions on*, vol. 53, no. 9, pp. 1979–1993, Sep. 2006.

[96] G. A. F. Seber and C. J. Wild, *Nonlinear regression.* New York: Wiley-IEEE, 2003.

[97] B. A. Freese and G. L. Buller, "A method for extracting SPICE2 junction capacitance parameters from measured data," *Electron Device Letters, IEEE*, vol. 5, no. 7, pp. 261–262, Jul. 1984.

[98] P. Ma, M. Linder, M. Sanden, S.-L. Zhang, M. Östling, and M.-C. F. Chang, "An analytical model for space-charge region capacitance based on practical doping profiles under any bias conditions," *Solid-State Electronics*, vol. 45, no. 1, pp. 159–167, Jan. 2001.

[99] H. Ding, J. J. Liou, C. R. Cirba, and K. Green, "An improved junction capacitance model for junction field-effect transistors," *Solid-State Electronics*, vol. 50, no. 7-8, pp. 1395–1399, 2006.

[100] S.-M. Kang and Y. Leblebici, *CMOS digital integrated circuits: analysis and design.* Boston, MA: McGraw-Hill, 2003.

[101] S. M. Sze, *Physics of semiconductor devices*, 2nd ed. New York: John Wiley & Sons, 1981.

Summary

Advanced Breakdown Modeling for Solid-State Circuit Design

by Vladimir Milovanović

Modeling of the effects occurring outside the usual region of application of semiconductor devices is becoming more important with increasing demands set upon electronic systems for simultaneous speed and output power. Analog integrated circuit designers are forced to enter regimes of transistor operation that are close to or within the device breakdown. They use compact models that describe device behavior in an efficient way to predict a designed circuit performance. Using modern heterojunction bipolar transistors with superb maximum unity current gain and maximum unity power gain frequencies, necessarily brings with it ever lower breakdown voltages. Impact ionization that causes avalanche multiplication has a profound impact on power amplifiers and plays a dominant role in the region of high output voltages, necessary for driving antennas of modern (ultra)wideband wireless systems.

On the other hand, digital circuit designs mostly suffer from high transistor leakage current that in the state of the art digital solutions takes up significant portion of the total power dissipation of a digital system. Therefore, it is of essence for digital integrated circuit designers to posses an accurate prediction of the leakages so that they may continue to grasp benefits of transistor downscaling.

In this thesis, starting from impact ionization, firstly, physics behind this phenomenon is studied. Frequency limitations of avalanche models are analytically derived in Chapter 2. A derivation is followed by the description of usual approaches for addressing impact ionization effects in semiconductor devices. Emphasized is the most frequently used, impact ionization rates approximation. The last part of the chapter is reserved for compact modeling of avalanche multiplication in semiconductor devices. This chapter presents a foundation for the two chapters that follow.

Chapter 3 focuses on quasidistributed bipolar transistor model reduction techniques. This model type is employed to describe complex multidimensional vertical current pinch-in effects that may occur in transistors biased within the avalanche region. A simplification method for the model is introduced, based on an implementation of bilinear approximation. Excellent matches between the original and reduced model are obtained. The model complexity is reduced from quadratic to linear dependency on size, nevertheless, the speed gain is not that dramatic.

Implications of impact ionization on bipolar transistors in terms of working in the small alternating signal environment are explored next. Specifically, in such cases avalanche characterization is important in order to proceed with deeper analysis. Chapter 4 gives a derivation proof that avalanche in the small signal drive conditions may be studied by observing the real parts of admittance parameters when transistor is viewed through its two-port network representation. Addressed are the needs for an accurate modeling of such regimes. Repercussions of avalanche on some important intrinsic active device properties from circuit design prospective are discussed in general. Collapse of unilateral power gain and increase of transistor stability are demonstrated and physically explained through the concept of intrinsic avalanche-induced negative feedback. The frequency above which avalanche effects in small signal conditions can be neglected is identified.

A description of a novel model for the band-to-band tunneling current in p-n junctions is shown in Chapter 5. The presented work consists of the model physical foundations, implementation and finally its verification against state of the art industrial and modern in-house device measurements. The developed tunneling breakdown model is fully physics-based and may be used both in bipolar as well as in field-effect compact transistor models. It is smooth in a mathematical sense on a whole real domain, thus escaping any potential solver convergence problems. The derived model features increased efficiency without compromising accuracy since it is not evaluated in the forward bias regime where the Zener tunneling current identically equals zero. Innovative parametrization of the model equation (in a statistical sense) drastically reduces the influence of randomness inevitably present in the measured data on which parameters are estimated, on dispersion of the extracted parameter values. As a consequence scaling over geometry and temperature is greatly improved.

Parameter extraction techniques in compact modeling in general have a crucial role. However, if several extraction methodologies for estimation of certain model parameter(s) exist, it is not trivial to select the best, that is, the preferred one. It is even unclear how “the best” strategy should be defined. Chapter 6 is devoted to this important topic, namely the analysis of parameter extraction strategies and parameter optimization. Since this thesis concentrates on modeling of breakdown phenomena that are driven by the electric field within the p-n junction’s space charge region, accent is drawn to the p-n junction parameters and their estimation methodologies. More precisely, obtaining parameters of the depletion capacitance and ideal diode current compact model parameters is covered in detail. The estimation strategies are compared in statistical terms which provide an insight in how the two, or more, can be assessed and compared, and which one would be more suitable for use in practice. In particular, it is demonstrated that it is much favorable to extract parameters simultaneously with their (temperature) scaling parameter rather than separately. Additionally, an approach to assess statistical properties of an arbitrary parameter extraction strategy and demonstrate the merits of such assessments is presented.

A collection of the main conclusions of the thesis is deferred to Chapter 7. It also provides the reader with several recommendations for future work.

Samenvatting

Geavanceerde Modellering van Doorslag voor het ontwerpen van halfgeleider circuits

door Vladimir Milovanović

Het modelleren van effecten die plaatsvinden buiten het gebruikelijke toepassingsgebied van halfgeleidercomponenten wordt belangrijker met de steeds hogere eisen gesteld aan elektronische systemen voor tegelijkertijd snelheid en uitgangsvermogen. Ontwerpers van analoge geïntegreerde circuits worden gedwongen om werkgebieden van transistoren te gebruiken die dichtbij of binnen het doorslaggebied van transistoren vallen. Zij gebruiken compacte modellen die het componentgedrag op een efficiënte manier beschrijven om de prestaties van ontworpen circuits te voorspellen.

Het gebruik van moderne heterojunctie bipolaire transistoren met uitmuntende maximale afsnijfrequenties brengt noodzakelijk steeds lagere doorslagspanningen met zich mee. Impact-ionisatie die lawine vermenigvuldiging veroorzaakt heeft een diepgaand effect op vermogensversterkers en speelt een dominante rol in het gebied van hoge uitgangsspanningen, noodzakelijk voor het aansturen van antennes van moderne (ultra)breedbandige draadloze systemen.

Anderzijds hebben ontworpen digitale circuits te lijden van voornamelijk hoge transistor lekstromen, die in de *state of the art* digitale oplossingen verantwoordelijk zijn voor een significant deel van de totale vermogensdisipatie van het digitale systeem. Daarom is het essentieel voor ontwerpers van digitale geïntegreerde circuits om te beschikken over een nauwkeurige voorspelling van de lekstromen zodat men kan blijven profiteren van de voordelen van transistorverkleining.

In dit proefschrift wordt eerst de fysica achter het fenomeen impact-ionisatie bestudeerd. Frequentielimitaties van lawinemodellen worden analytisch afgeleid in Hoofdstuk 2. De afleiding wordt gevolgd door de beschrijving van gebruikelijke benaderingen die worden toegepast bij de praktische modellering van impact-ionisatie effecten in halfgeleider componenten. Benadrukt wordt de meest frequent gebruikte benadering: de impact-ionisatietempo benadering. Het laatste deel van het hoofdstuk is gereserveerd voor compacte modellering van lawinevermenigvuldiging in halfgeleider componenten. Dit hoofdstuk presenteert de basis voor de twee hoofdstukken die volgen.

Hoofdstuk 3 richt zich op quasi-gedistribueerde bipolaire transistor modelreductie technieken. Deze technieken worden hier gebruikt om complexe multidimensionale verticale stroom *pinch-in* effecten die kunnen optreden bij transistoren ingesteld in het lawinegebied te beschrijven. Een vereenvoudigingmethode voor een model wordt geïntroduceerd, gebaseerd op een implementatie van een bilineaire benadering. Uitstekende overeenkomsten tussen het originele en gereduceerde model zijn behaald. De modelcomplexiteit is gereduceerd van kwadratische tot lineaire afhankelijkheid van grootte, nochtans is de snelheidswinst niet zo dramatisch.

Implicaties van impact-ionisatie op bipolaire transistoren met betrekking tot de werking in de kleine alternerende signaal context worden vervolgens verkend, met name in die gevallen waarin lawine karakterisatie belangrijk is. Hierna volgt een diepere analyse. Hoofdstuk 4 geeft een afgeleid bewijs dat lawinevorming in de klein-signaal aanstuurvoorwaarden bestudeerd kan worden door de reële delen van de admittantie parameters te bekijken, wanneer de transistor wordt bekeken in zijn tweeport netwerk representatie. De behandeling richt zich naar de behoeften aan een nauwkeurige modellering van zulke regimes. Gevolgen van lawinevorming op sommige belangrijke intrinsieke actieve componenteigenschappen, die belangrijk zijn in de context van een circuitontwerp, worden in het algemeen behandeld. Het instorten van unilaterale vermogensversterking en het toenemen van transistor stabiliteit worden gedemonstreerd en fysisch verklaard door het concept van intrinsieke lawinegeïnduceerde negatieve terugkoppeling. De frequentie waarboven lawine effecten in kleinsignaal condities kunnen worden verwaarloosd is geïdentificeerd.

Een beschrijving van een nieuw model voor de band-naar-band tunnelstroom in p-n juncties wordt getoond in Hoofdstuk 5. Het gepresenteerde werk bestaat uit de fysische basis van het model, de implementatie en tenslotte verificatie aan de hand van metingen aan *state of the art* industriële en moderne academische componenten. Het ontwikkelde tunneldoorslagmodel is volledig gebaseerd op fysica en mag gebruikt worden zowel voor compacte modellen van bipolaire transistoren als voor modellen van veldeffecttransistoren. Het is glad in wiskundige zin op het gehele heel reëele domein, en voorkomt zodoende potentiële numerieke convergentieproblemen. Het afgeleide model brengt aldus een toegenomen efficiëntie zonder een afbreuk te doen aan de accuraatheid, aangezien het niet wordt geëvalueerd in de doorlaatrichting waar de Zener tunnelstroom gelijk is aan nul. Innovatieve parametrisatie van de modelvergelijking (in statistische zin) reduceert drastisch de doorwerking van stochastische fouten, die onvermijdelijke aanwezig zijn in de gemeten data waaruit de parameters worden geschat, in de dispersie van de geëxtraheerde parameterwaarden. Bijgevolg is de schaling over geometrie en temperatuur sterk verbeterd.

Parameter extractie technieken in compacte modellering hebben over het algemeen een cruciale rol. Echter, als meerdere extractie methodologieën voor de schatting van bepaalde modelparameters bestaan is het niet triviaal om de beste te kiezen, dat wil zeggen, degene waar de voorkeur naar uitgaat. Het is zelfs onduidelijk hoe "de beste" strategie gedefinieerd dient te worden. Hoofdstuk 6 is gewijd aan dit belangrijke onderwerp, namelijk de analyse van parameterextractiestrategieën en parameteroptimalisatie. Aangezien dit proefschrift zich richt op de modellering van doorslagfenomenen die worden gestuurd door het elektrische veld binnen de ladingsdragervrije zone van de p-n junctie wordt het accent gelegd op de p-n -junctieparameters en hun schattingsmethodologieën. Meer precies: het verkrijgen van parameters van de depletiecapaciteit en van de compacte modelparameters voor de ideale diodestroom, wordt in detail beschreven. De schattingstrategieën zijn vergeleken in statistische termen, die een inzicht geven in hoe twee methoden, of meerdere, kunnen worden beoordeeld en vergeleken, om te zien welke meer geschikt is voor gebruik in de praktijk. In het bijzonder is het aangetoond dat het veel gunstiger is om de parameters simultaan met hun (temperatuur-) schalingsparameters te extraheren in plaats van apart. Bovendien worden een aanpak om de statistische eigenschappen van een arbitraire parameterextractiestrategie te beoordelen en de verdiensten van een dergelijke beoordeling, gepresenteerd.

Een overzicht van de hoofdconclusies van dit proefschrift wordt uitgesteld tot Hoofdstuk 7. Deze biedt de lezer ook verscheidene aanbevelingen voor toekomstig werk.

List of publications

Journal papers

1. **V. Milovanović** and R. van der Toorn, “A Novel Physics-Based Compact Model of Band-to-Band Tunneling Current in p-n Junctions”, *IEEE Transactions on Electron Devices*, July 2010, volume 57, issue 7, pages 1583-1589

2. **V. Milovanović** and R. van der Toorn, “Impact of Parameter Extraction Methodology on Variances of Extracted Parameter Values”, *Solid-State Electronics*, June 2010, volume 54, issue 6, pages 665-670

3. **V. Milovanović**, R. van der Toorn and R. Pijper, “RF Small Signal Avalanche for Bipolar Transistor Circuit Design: Characterization, Modeling and Repercussions”, *Microelectronics Reliability*, submitted for publication

4. M. Popadić, **V. Milovanović**, C. Xu, F. Sarubbi and L. K. Nanver, “C–V Profiling of Ultrashallow Junctions using Step-Like Background Profiles”, *Solid-State Electronics*, special issue, in press, available online

Conference proceedings

1. **V. Milovanović** and R. van der Toorn, "On p-n Junction Depletion Capacitance Parameter Extraction Strategies", *Proceedings of 27th International Conference on Microelectronics Proceedings, MIEL 2010*, Niš, Serbia, 16-19 May 2010, pages 87 - 90

2. **V. Milovanović**, R. van der Toorn, P. Humphries, D. P. Vidal and A. Vafanejad, "Compact Model of Zener Tunneling Current in Bipolar Transistors featuring a Smooth Transition to Zero Forward Bias Current", *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting, BCTM 2009*, Capri, Italy, 12-14 October 2009, pages 99-102

3. **V. Milovanović** and R. van der Toorn, "RF Small Signal Avalanche Characterization and Repercussions on Bipolar Transistor Circuit Design", *Proceedings of the IEEE EUROCON 2009*, Saint Petersburg, Russia (Russian Federation), 18-23 May 2009, pages 230-233

4. **V. Milovanović** and S. Mijalković, "An Efficient Sectionalized Modeling Approach for Introduction of Distributed Avalanche Effects in Bipolar Circuit Design", *Technical Proceedings of the NSTI Nanotechnology Conference, Nanotech 2007*, Santa Clara, California, 20-24 May 2007, volume 3, pages 658-661

Workshops

1. **V. Milovanović** and R. van der Toorn, "Statistical Analysis of Diode Ideal Current Parameter Extraction Procedures", *Proceedings of 12th Annual Workshop on Semiconductor Advances for Future Electronics and Sensors, SAFE 2009*, Veldhoven, the Netherlands, 26-27 November 2009, pages 87-90

2. **V. Milovanović** and R. van der Toorn, "Modeling Advanced Avalanche Effects for Bipolar Transistor Circuit Design", *Proceedings of 11th Annual Workshop on Semiconductor Advances for Future Electronics and Sensors, SAFE 2008*, Veldhoven, the Netherlands, 27-28 November 2008, pages 569-572

Given publication list is from the period of last four years.

Acknowledgments

Most probably everyone would tell the same about his own experience, but I really feel that my PhD adventure was a bit different than the usual one. I started well despite the fact that I came from completely different field. I managed to enter the new subject pretty quickly. Then a lot of things had changed. Not that all of them had an impact on me, but my initial daily supervisor left, the group I was in broke apart, even the institute I was working in changed its name, to name a few changes. I have not published anything for almost two years. I had my downfalls on the personal side as well. Many even started to doubt in me, even myself at one point. However, I was able to raise above the situation, be persistent, continue trying even when the things did not go in the wanted direction. No way I could do this alone. I am as happy as privileged to be in a position to express my gratitude to all of you who had helped me during the period of the last four years. Without you, even if I would somehow reach the end, I would simply not feel the way I feel right now.

First of all, I would like to thank Lis Nanver for finding enough good reasons to accept me as her promovendus after the interview I had with her more than half a decade ago. After that interview work prospectives did not look bright, but nevertheless things worked out that way that eventually she ended up being my promotor. Yes, I still remember my first workday in Delft, entering her office and after, "Hi, Lis!", from my side receiving a question, "Sorry, who are you?", from the other side. And, yes truly, throughout the time we did not have that much contact with each other. Only in the occasional evaluation talks, ECTM excursion or SAFE dinners we had a real chance to talk. Nevertheless, all the time simply knowing that if it really becomes tough in certain aspects I could come and solve that out with you, meant a lot to me. Only at the very end, with signing PROM forms from 1 to $+\infty$ and back, reviewing propositions, etc., I saw that my beliefs were right, that when I needed you, you were there. Thanks for that and for your special touch in my ten propositions.

Most probably, I am a PhD student of Lis with whom she had spent the least amount of time, but there is the guy on the other extreme point. He can definitely tell better about that experience, but I am sure that being a *daily* supervisor to a person that cannot stand himself every single day of the year is not a trivial job. Cool as usual and armed with (almost) infinite patience[§] whatsoever, I may say from my side that he did a pretty good job. I would like to thank my copromotor Ramses van der Toorn for many things but first and foremost of all for giving me as much freedom in the work as I wanted and for taking care about certain things that went beyond his job description. Pretty atypical for a Dutch guy, but maybe it is up to my own false perceptions as I met too many atypical Dutch persons. Also, organized as you are, you were a perfect counterbalance to my chaotic nature. Large part of this thesis would not look as it looks now if there was not for you.

[§] I feel relevant to declare this because of the adjective *daily*. As this adjective is missing in front of Lis' name, her patience limits, to this end, remain unexplored (and unknown at least to me).

Third atypical Dutch person that I would like to thank is my ex-supervisor and never-meant-to-be-my-copromotor Slobodan Mijalković. He is the single most responsible person for the fact I started my PhD studies in Delft in the first place. Since the very beginning on the interview I had, when after my lapsus a lot of dust went up in the air, he was protecting me as no one. Afterward, when he knew he was leaving the faculty he set everything in my favor. Even when he left, rarely, but regularly, we met and discussed how the things were going for me. Slobo, thank you for that and for everything else outside the work activities.

One other guy (also from these Acknowledgments) said that (I paraphrase) work is the best when mutual interests exist. I reckon, not much is more true than this. This thesis is, I guess, the best way to thank all off you. It certainly is the biggest thing I got from you, because without you, this very thesis would not be possible. Thank you once again.

Talking about it, I express my deepest gratitude to all committee members for accepting to judge my thesis and devoting their time to be part of the doctoral graduation ceremony.

Staying within Dimes, I would like to express my special thanks to Pasqualina Sarro for her unmatched ability to convey optimism to others, that is, in other words, to put others to work. From the experience of other guys I was never afraid that I might be temporarily unemployed. I do not know why am I putting this sentence in this place, but nonetheless, luckily, I had alternatives. Without Lina, ECTM would not look as it looks now.

I would like to acknowledge Leo de Vreede for taking care of me when I was without supervisor, for several nice practical suggestions and for directness in his every approach.

I am grateful to Peter Swart for acquainting me with everything around the DC measurement room and Atef Akhnoukh for teaching me everything I know (and more) about RF measurements, and not giving up on me even when I was breaking needles and probes.

Also thank Henk van Zeijl for inviting me to my first PhD defense and making, so spectacular jokes that others were even afraid I will not understand them as such.

I have not worked in the cleanroom so it remains a mystery to me what is happening down there. Notwithstanding, I would like to say thanks to every member of the Dimes IC processing group for coming to my coffee pluses and friendly chats in the coffee corner.

I am very grateful to all the smiley secretaries that worked and work within Dimes for helping me to find a cut and escape an infinite loop of Dutch bureaucracy.

I feel that without Rino Martillia Dimes simply is not the same place it use to be. Rino thanks for making a positive and cheerful difference wherever you might appear.

I cannot find the words which could express my gratitude to Silvana Milosavljević (and Slobodan). Since the very beginning you have accepted me as someone you have already known for years, inviting me to every occasion you might celebrate or just hang around, and simply for being a truthful friend(s) of mine in every sense of that expression.

Not everything is about work in Dimes. Something is in just pure relaxation as well. I was honored to have a workplace in such an extraordinary multicultural environment.

I must first thank the “old school” fellow comrades, now mostly graduated PhDs:

- Luigi La Spina for being ready to give me some smart advice no matter what about and for possessing such an unprecedented sense of humor that shined in full glory during the phenomenal nationality quiz in the last year’s ECTM annual review meeting.
- Gianpaolo Lorito for his often accurate but always very interesting verbal connotations.
- Franceco Sarubbi for his sincereness and words of wisdom whenever I may asked for.
- Yann Civale times positively surprising me for many on how much he was capable of.
- Koen Buisman for showing to all of us what a unit of devotion and interest, not to work, rather a research or science, would be and for being around since my interview.
- Theodoros Zoumpoulidis for setting up so tight PhD schedule standards that you feel relaxed even when you should work nonstop day and night in order to finish on time.
- Jia Wei for at least to some extent compensating this bad habit in the photo finish.

I would also like to acknowledge the colleagues that have not been in Dimes when I arrived. I can tell that every single one of you brought something new, that is, in particular I thank:

- Cleber Biasotto for being the first to congratulate when there where reasons to do so.
- Fabio Santagata for being such a good Pulcinella (or should I write it originally as Pulecenella?) even in everyday activities, that better one I could not even think of.
- Daniel Tajari Mofrad for curious off the record conversations when we had time for.
- Sten Vollebregt for readily translating in Dutch thesis summary and propositions.
- Parastoo Maleki for breaking seldom silence in our office with yet another question.
- Aslıhan Arslan for bringing with her some positive change within the last few months.
- Daniel Vidal, mostly, but not limited to, that night of "fear and loathing in Naples", and for repeatedly being able to question my thoughts and to help me when necessary.
- Francesco Vitale for pushing my philosophical skills to the very limits on various topics such as "what is the meaning of the PhD studies and research in general".

Very special acknowledgment goes to the best office mates around (past and current), Lei Gu, Amir Sammak and Ana da Silva, whom I would not trade for anyone else.

Going a bit further away from the faculty community I would like to start with Dejan Vlašić and Biljana Ćamilović, my friends from Rotterdam and (of course originally Kruševac) I could rely on from the very beginnings of my long journey. It is impossible to thank them for their continuous encouragement and firm support throughout these years.

Next I would like to say a few words about every member of my Serbian(-speaking) community with whom I never felt so far away from home as objectively I was. In particular I want to thank Filip Miletić for his unique repartees in lucrative talks during joint lunches in Aula until he left, Nebojša Nenadović for actually being the one who acquainted me with Delft and for his countless invitations to visit him and Nela in their new home, Aleksandar Borisavljević on numberless bets and quotes to them and his Veronica for spiced irregularities in our perfectly-speaking community, Darko and Jasmina Simonović for their balanced attitudes in our converse, one of my neighbors Ivan Lazić for always being ready to pick up a guitar, sing and animated us, Stevan Nađ-Perge for fruitful debates ranging from physics to politics, Dubravka Aranđelović for her uncountable offers for a meal and pleasant company in Rijswijk, Agata Šakić and Luigi Mele for being charming and amusing companions in any respect, Steva and Maja Rudinac for their careful recommendations and proposals of various kind, Mihajlo Obućina for enjoyable moments in Delft's cafés and other opportunities, Miloš Aćanski for demonstrating to me in practice where is my place in Worms, Miloš Popadić for numerous technical and philosophical debates and to both him and his lovely wife Jelena for their friendship and for being ready to help me in every aspect, then of course my second neighbor and "zemljak" Marko Mihailović for being in charge of organization of anything that could be organized (except for himself) where I was always welcome to join and for assisting me in everything I needed assistance regardless of the time and place, and my "imenjak" Vladimir Jovanović for his unmatched wideness and deepness that he was willing to share in every talk no matter what the subject was. I was delighted to have you all here.

Further, I wish to acknowledge all of my Russian(-speaking) friends, close to me or far away, that I was lucky to meet and have contact with. Also in particular I would like to thank: Olga Vladimirovna (family name I do not know anymore) for those magic half a year of your life that you unconditionally devoted to us, it meant a lot to me back then, Maria Rudneva for being willing to talk to me when I was everything but talking, to Vadim Sidorkin for being a perfect compatriot of the future Slavic Union, Yevgeniy Pivak for always being able to drink one more. Very special acknowledgment goes to my best friend and undoubtedly the most important person I met during my recent endeavor. Slava, thanks for being a perfect teacher of Russian, even better life philosopher and so compatible soul-mate. If I would even mention other things you did for me, these lines would lose their true meaning.

I acknowledge all the professors and assistants of my *alma mater* that gave me such a solid theoretical background. However, among all of them I must emphasize my gratefulness to Predrag Pejović and Lazar Saranovac who boosted my motivation half a year ago with one very special two-hour talk. Without that discussion, this thesis would not have been finished today. I would like also to thank Radivoje Đurić for more or less regular meetings for a drink or two in which he was unselfishly sharing his views on life, love and electronics.

Acknowledging all the friends that were not physically present near by is distinctive from the reset. The fact that even though they have been far away and still managed to find the time to send, if not more then just, a greeting, had a tremendous meaning to me. Particularly, I thank: Nikola Staničić – Nikica for cherishing such an atypical nature and spirit with whom I could always escape everyday life for a while, for his endless talks over the Internet and for being the first one to visit me here, Novak Novaković – Novačić for so honest boasts that gave all of us around the opportunity to reply, Milena Novaković for always providing a backup of anything, Nikola Todorović for his candid indications and "no problem" answers on every problematic question, Boris Jovanović for so original array of jests he would exhibit and for long lasting friendliness we have, Srđan Milić for being such a "nepopravljiva baraba", Nebojša Glišić for always being around for a hang out when I am in my hometown, Ivana Ilić for her proper dose of irony whatever we were talking about, Anđelka Brkić for forgiving me everything as nothing happened without even me knowing that something happened, Ivana Milenković for words and words of advice and anything else that I might needed, that went on to be true even when I deeply doubted in them, well with the exception for political ones, which were, are and will remain to be just wrong until she changes them, Jelena Nenadović for her lucidness and rare but exhaustive emotional eruptions and long night-morning talks, Jovana Jović and Milena Stojković, whom I have been so fortunate to meet, for their pleasing and refreshing kindness with which they approached every time, Marina Plavšić for always making plans to see each other more often and Marko Ćirković – Dadica, for always being ready to change his mind in my favor, for being faithful and frank even when it was against his own interests in the short run. Special thank goes to all of the members of the "Elektronika i Prijatelji" group for entertaining discussions on every topic but electronics. I would like to thank those few girls I have an NDA with for their rare and intensive (as rare and as intensive one may imagine) encouragement.

At the end, I would like to thank members of my family. To them I devote this thesis.

На крају ове тезе, желео бих да се захвалим свој својој породици. Појединачно, почео бих од најмлађих, а опет довољно старих да су за претходне четири године, колико сам у Холандији, оставили довољно трага на мене. И тако, хвала: Андреји и Ивану за изузетно друштво сваки пут кад бих долазио у Србију, ту деценијску разлику у годинама са вама скоро никада нисам осећао, те Ацку и Марку и Александри и Аци, које истина нисам тако често виђао, али кад год би се срели задовољство је било обострано и нескривено. Такође, захваљујем својим сестрама Ивани и Тањи које су биле ту у најтежим тренуцима, сестрама Весни и Лели, коју опет видех само једном на кратко за све претходно време, браћи Дарку и Деји али и Гаги и Жики за помоћ сваки пут кад би ми затребала. Захвалнице иду и нашим кумовима, које исто тако сматрам делом породице, на њиховим срдачним речима подршке у сваком међусобном разговору. Хвала увек драгим течи Гоји, ујка Мићору и ујни Радици и нани Нати, чика Милету и стрина Вери који су ме изнова богато угошћавали кад год би отишао у Придворице, Београд или пак Милошевац. Тетке какве ја имам, ретко се срећу. Једно посебно хвала тетка Мари, тетка Љиљи и тетка Миланки које на мене гледају на истински мајчински начин, али и да знају да су та осећања узајамна. Немерљива је моја жал што моја баба Боса и мој деда Мића нису доживели да, ако не сами, онда да им барем неко други прочита ове речи. С друге стране срећан сам што ће бар моја Бикица то дочекати. На самом концу, благодаран сам својој мајци Светлани за безрезервну љубав коју ми је пружала цело ово време. Хвала вам најлепше свима. *Вама* посвећујем ову тезу.

Delft, 28 June 2010 / Делфт, 28. јун 2010. год.

About the author

Vladimir (Milan) Milovanović (Serbian: Владимир (Милан) Миловановић) was born in Smederevska Palanka, Serbia, on 2 October 1981. He received the Dipl.-Ing. from the Faculty of Electrical Engineering, University of Belgrade, Serbia, in July 2005, graduating from the Department of Electronics with a novel FPGA implementations of bioinformatic algorithms. During the studies, in 2004, he was a summer student researcher in the Signal and Image Processing Laboratory, Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa. The work done at the Department of Electronic Engineering (DIE), Faculty of Telecommunication Engineering (ETSIT), Technical University of Madrid (Universidad Politécnica de Madrid, UPM), Spain, in the spring of 2005 led to his Master's thesis.

Afterwards, for half a year he worked as a Principal Engineer in the area of high speed digital signal processing solutions at Signum Concepts, Inc., Belgrade office.

In March 2006 he joined the Department of Microelectronics and Computer Engineering, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology (Technische Universiteit Delft), Holland, where he worked toward the PhD degree within the Delft Institute of Microsystems and Nanoelectronics (DIMES). His research was focused on compact physics-based modeling of breakdown mechanisms in semiconductor devices for use in integrated circuit design.