Development and Evaluation of Novel Chip Modification- and Analysis-Techniques, based on Backside Focused Ion Beam Preparation

vorgelegt von

Dipl.-Ing.
Rudolf Schlangen
aus Neuss

Von der Fakultät IV -Elektrotechnik und Informatik
Institut für Hochfrequenz- und Halbleitersystemtechnologie
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. K. Petermann
Berichter: Prof. Dr.-Ing. C. Boit
Berichter: Prof. J. Melngailis

Tag der wissenschaftlichen Aussprache: 16.12.2009

Berlin 2010

D 83

Erklärung:

Ich versichere an Eides statt, dass ich die Dissertation selbstständig verfasst habe. Die benutzten Hilfsmittel und Quellen sind in der Arbeit vollständig angegeben.

Zusammenfassung:

Im Entwicklungsprozess moderner integrierter Schaltkreise hat sich seit den späten Achtzigerjahren der Bereich der ionenstrahl-basierten Schaltungsmodifikation fest etabliert. Mittels „Focused Ion Beam“ (FIB) Gerätschaften ist es hierbei möglich, Fehler in den Verdrahtungsebenen neuer Schaltkreise zu reparieren, und somit durch weitere Messungen alle vorhandenen Fehler, und deren Lösungen, bereits mit den ersten hergestellten Schaltkreisen ausfindig zu machen. Mit der stetigen Weiterentwicklung der Halbleitertechnologie, der steigenden Integrationsdichte, der wachsende Anzahl an Verdrahtungsebenen und der Anwendung immer zahlreicherer Materialien im Herstellungsprozess der Schaltkreise stoßen jedoch die etablierten Vorgehensweisen mehr und mehr an ihre Grenzen. Die Schaltungsmodifikation durch die Schaltkreisrückseite (Siliziumträgermaterial) bietet hier eine sehr vorteilhafte Alternative, da der Zugang zu den unteren Verdrahtungsebenen so enorm erleichtert wird. Der Schlüssel zu erfolgreichen rückseitigen Modifikationen ist dabei der kontrollierte- und planparallele Abtrag des Siliziumgrundmaterials im Zielgebiet.

Den Kern dieser Arbeit bildet die Untersuchung der Auswirkungen dieses rückseitigen FIB Prozesses auf die statischen und dynamischen Leistungsparameter von Bauelementen und Schaltkreisen, produziert in einer 120 nm Technologie. Während sich die statischen Bauteil-parameter um ca. 10% verschlechterten, zeigten die dynamischen Schaltungseigenschaften eine Verbesserung um 10% bis 60%, in einer komplexen Abhängigkeit von der verbleibenden Schaltungsrestdicke, der Versorgungsspannung und der initialen Schaltungseigenschaften. Mithilfe von physikalischen Simulationen konnte ein umfassendes theoretisches Modell abgeleitet werden, welches die gemessenen Veränderungen einzelnen Effekten zuordnen lässt und somit Vorhersagen für die Auswirkung auf andere Schaltkreise (auch anderer Technologien) zulässt.

Dies ermöglicht die gezielte Anwendung des FIB Prozesses um einzelne Schaltungsteile nachträglich zu beschleunigen und somit Laufzeitproblemen zu begegnen, für die es bislang kaum adäquate Lösungen gab, die aber mit steigender Signalgeschwindigkeit immer häufiger zu Problemen führen.

Des Weiteren wurden alternative Kontaktierungsmethoden entwickelt, charakterisiert und erprobt, die klare Vorteile bezüglich ihrer Anwendbarkeit und ihrer Kontakteigenschaften gegenüber den herkömmlichen Methoden zeigen.

Summary:

Ion beam based circuit modifications became a standard technique, applied during chip development since the late 1980s. Dedicated focused ion beam (FIB) tools allow repairing faults by rerouting the metal interconnects of a given circuitry. Due to the progress in semiconductor technology, with its increasing integration density, high number of metal layers and the usage of an ever increasing number of materials, the established FIB techniques are more and more facing their limitations. Through Silicon-, or backside circuit edit (CE) is an advantageous alternative, offering better access to the lower interconnect levels. The key for successful backside CE is the controlled and co-planar removal of the bulk Si in the area of interest.

The main part of this work consists of an in-depth invasiveness study, where the impact of the FIB backside thinning onto static and dynamic circuitry performance is evaluated by use of 120 nm technology single FETs and ring oscillators. The static device performance decreased by ≈10% in contrast to which the dynamic circuitry performance increased by 10% up to 60%, showing a complex dependence on the remaining Si thickness, the core supply voltage and the underlying circuitry design. Using physical device simulations to reproduce and complement the experimental results allowed deriving a complete model, which links the performance alterations to certain physical effects, allowing forecasting the impact of the proposed procedure on different circuitries and technologies.

Based on this understanding, the FIB process can now be utilized to modify the chip internal timing in a controlled and non-destructive way, which is often not possible with today's established methodologies, but highly desired since these timing marginalities (or soft-fails) become a predominant limitation for recent and future high speed circuitries.

Furthermore, alternative contact methodologies were developed, characterized and successfully applied, showing clear advantages in terms of their contact properties and general applicability compared to the established techniques.

“Everything should be made as simple as possible,
but not one bit simpler.”

“Man sollte alles so einfach wie möglich sehen,
aber auch nicht einfacher.”

Albert Einstein
(1879 - 1955)

Table of Content

List of Selected Symbols

Symbol	Description	Unit
$\overline{\varphi_b}$	average floating body potential	V
τ	average inverter delay	s
φ_b	floating body potential	V
ΔI_D	normalized drain current reduction	%
$\Delta I_{D,nT}$	non-thermal normalized drain current reduction	%
τ_n	n-FET delay	s
τ_{osc}	oscillation period (RO)	s
τ_p	p-FET delay	s
C’	capacitance per unit area	Fm^{-2}
$C_{de\ ch}$	channel depletion capacitance	F
C_{drain}	FET drain junction capacitance	F
C_{GB}	gate to body capacitance	F
C_{in}	input capacitance of CMOS gate	F
C_L	external (FET extrinsic) load capacitance	F
$C_{L\ sim}$	load capacitance for simulation	F
C_{out}	output capacitance of CMOS gate	F
C_{ov}	gate to source/drain overlap capacitance	F
C_{ox}	gate-oxide capacitance	F
C_{source}	FET source junction capacitance	F
E	energy (in band-diagram)	eV
E_c	conduction band energy	eV
E_f	Fermi-level	eV
E_v	valance band energy	eV
FO	fan out	-
$I_{DB\ rev.}$	drain to well diode reverse current	A
$L = L_{eff}$	effective gate-length	m
L_d	drain-length	m
L_{diff}	diffusion-length	m
L_{gate}	physical gate length (poly structure)	m

Symbol	Description	Unit
L_{layout}	gate length in layout	m
m	body-effect coefficient	-
n	number of inverters of an RO	-
Q_i	inversion-layer charge	As
R'_{on}	FET on-resistance times unit width	Ωm
R_{leak}	symbol for FIB-induced leakage	Ω
R_{on}	FET on-resistance	Ω
R_{sw}	switching resistance	Ω
R_{th}	thermal resistance	cm^2 K/W
R_{well}	well-contact resistance	Ω
t_{fall}	linear signal fall time (input)	s
t_{ov}	gate to source/drain overlap	m
t_{ox}	gate oxide thickness	m
t_{rise}	linear signal rise time (input)	s
t_{Si}	remaining Silicon thickness	m
V_{bi}	build-in voltage	V
V_{dd}	CMOS supply voltage	V
V_{fb}	flat-band voltage	V
V_{in}	CMOS input voltage	V
V_{out}	CMOS output voltage	V
$V_{rev.}$	reverse-biased voltage across pn-junction	V
V_{ss}	CMOS ground potential	V
V_t	theoretical threshold voltage	V
$V_{t\ lin}$	linearly extrapolated V_t (with $V_{DS} = 50$ mV)	V
W	width of FET	m
w^0_d	depletion-region width	m
$w^0_{d\_m}$	maximum depletion-region width	m
W_g	Silicon bandgap energy	eV
ξ	electric field	V/m
ψ_B	volume/body potential	eV
ψ_S	surface-energy/band-bending	eV

1. Introduction

In today's highly engineered world the progress in many fields of research and development is directly linked to the improvements of the semiconductor industry, delivering ever increasing computation power at ever decreasing costs. In contrast to the early years of this industry, when most innovations originated from military projects, this development is now almost completely driven by private and globally operating companies focusing on the vast, yet still rapidly growing consumer market.

Especially for mass products like office PCs, dedicated gaming equipment and handheld devices the different companies work under the increasing pressure of competition; and the company with the fastest *time to marked* wins.

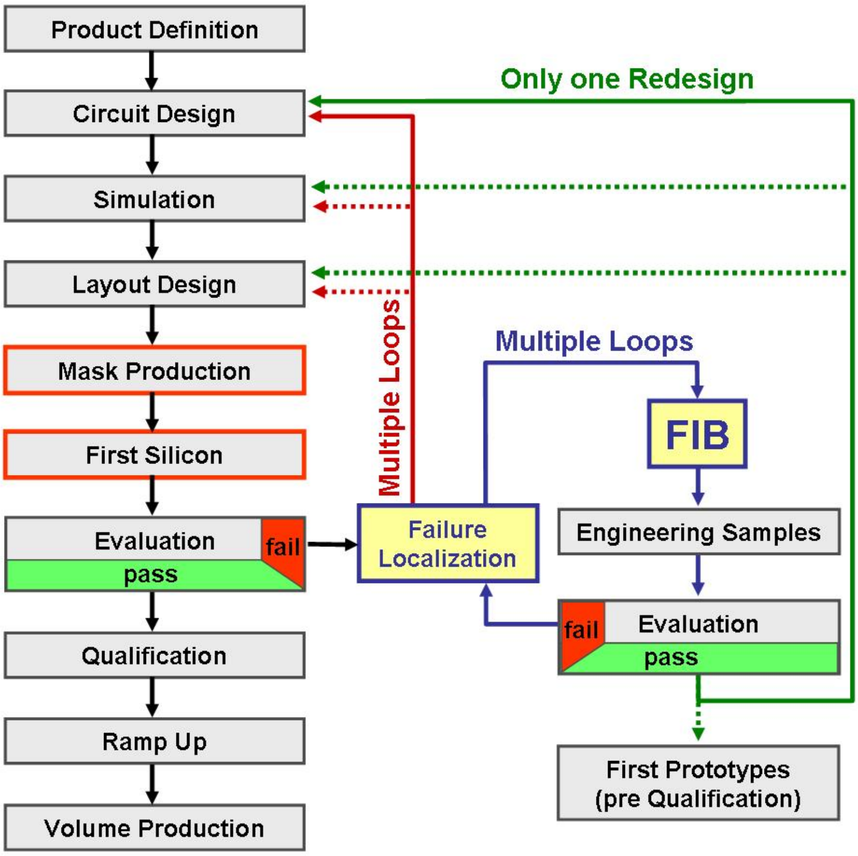

Figure 1-1: Chip development flow with and without the use of FIB

The typical development phases for a new semiconductor product are summarized in Figure 1-1. The definition of the desired product features is followed by the circuit design and simulation resulting in the generation of the physical layout. After that, the mask set can be produced, allowing the fabrication of the first silicon samples followed by their evaluation. In the likely event of facing marginalities or fails, the root causes have to be determined by means of failure localization. Before the use of focused ion beam (FIB) tools became part of the standard procedure in the early 1990s [1] & [2], the whole process would then have to be repeated (red loop). Besides the changes in circuit design and layout, this enforces the production of at least one new mask and the repetition of the backend fabrication process (may require the repetition of the full process if changes on the device layer were necessary). These two steps are both not only very time (1-3 month) and cost intensive, following this old way of chip development there is a high probability of finding additional fails after the first redesign, either because these fails were masked by the ones found earlier or as a result of altered conditions due to the design change. In that case, the whole procedure would have to be repeated again, losing additional time and money.

The use of FIB circuit edit (CE) allows repairing single engineering samples. Following the blue branch, the repaired samples can be evaluated again without any further delay except for the FIB modification time. Repeating this short loop can ensure that all fails and marginalities of the initial design can be found based on these first silicon samples. This strongly increases the confidence level for the final redesign and can drastically reduce the overall debug time and costs, hence also reduce the time to market. In addition, it may be possible to build first fully functional prototypes based on the FIB repaired chips, allowing for demonstration and/or pre-qualification with potential customers.

Today's FIB CE tools evolved from pure ion milling units with the introduction of process gasses enabling material selective etching and the deposition of insolating and conducting material in a direct writing manner [3] & [4]. But in comparison to the early years of FIB CE (1980s), when chips were fabricated with two or three non-planarized metal layers, recent technology generations demand by far more advanced FIB capabilities. Especially the tendency to incorporate an increasing number of metal layers in combination with low-K inter-metal dielectrics, flip-chip packages and system in package approaches drastically limit or fully block the access through the structured frontside [5].

These problems are also relevant for most failure localization techniques, which drove the development of backside analysis and stimulation tools, using the transparency of silicon for near-infrared (NIR) light [6], [7] & [8].

In terms of circuit edit, dedicated backside FIB tools were introduced allowing to modify the lower interconnect levels through the bulk silicon [9], [10] & [11].

But the NIR optical techniques as well as the established backside FIB procedures are facing their physical limits due to the increasing integration density and reduced feature sizes with recent and future technology generations [12] & [13].

The aim of this work is to expand the FIB CE and failure localization capabilities to cope with the increasing pace of chip development.

Several innovations will be presented and discussed; all based on a proposed backside FIB thinning procedure, reducing the thickness of the active FET layer to approximately 350 nm (referred to as ultra thin silicon - UtS).

The capability of CE could be expanded by the development of a novel contact methodology named contact to Silicide (CtS) which gives access to any node (or signal) on a given chip and allows for modifications on the transistor level for the first time.

UtS has also been successfully used as a new platform for probing, using old frontside electron beam probing systems to measure the desired timing information directly on the exposed transistors, showing superior lateral resolution and acquisition speed. Also, the resolution of optical techniques such as laser stimulation or laser voltage probing can benefit from the FIB thinning by using shorter wavelengths.

The central part of this work is an in-depth investigation of the FIB induced impact on the static and dynamic circuit performance caused by the UtS formation. Single FETs as well as differently designed ring oscillators of the 120 nm technology node (provided by Infineon Technology AG) were used for these experiments. As a result of the FIB thinning, the single FETs suffered from increased self-heating and channel mobility degradation whereby the dynamic performance of ring oscillators improved by 10% up to 60% depending on the ring design and core supply voltage.

A full model will be derived to explain the complex alteration of the FET intrinsic device physics allowing weighting the different identified contributors to the FIB induced delay reduction based on the help of static and dynamic physical device simulation, showing good agreement with the experimental data.

Based on the fundamental understanding of the complex origins of the speed gain, this technique can now be used to modify the chip's internal timing in a predictable and fully non-destructive way. This opens a whole new application field for backside FIB processing and offers highly desired additional solutions for dealing with soft-fails in recent and future technology generations where these timing issues are becoming the predominant limitation.

This thesis is structured as follows:

Chapter 2 begins with the introduction of FIB for circuit edit and discusses the costs and benefits of the backside CE compared to the frontside approach.

The utilized backside FIB tool with its special features and the established and later utilized backside CE procedures are described and discussed in chapter 3.

The next chapter (4) delivers the necessary semiconductor physics- and circuit design theory for the later discussion of the FIB induced performance alterations.

The utilized physical simulation environment (Synopsys® Sentaurus) is introduced in chapter 5 in combination with the adjustments of the implemented DC FET models.

The chapters 6 and 7 accommodate the main part of this work with an in-depth investigation of the invasiveness of the proposed FIB thinning procedure. Experiments on single n- and p-FETs are utilized in combination with intensive physical device simulations, allowing separating and weighting different degrading effects including increased self-heating, SOI related floating body effects, and finally the reduction of the carrier mobility in the FET channel in chapter 6.

The influence on dynamic circuit performance is evaluated in chapter 7. Combining experimental and simulation results yields a complete theoretical model with several contributing effects.

Based on the theoretical understanding, the FIB induced speed gain is discussed as a whole new application filed for trimming of chip internal timing conditions in chapter 8.

Chapter 9 presents UtS related innovations and improvements of backside CE. Especially one novel contact methodology called contact to Silicide (CtS) is discussed in more detail before the last chapter (10) evaluates the potential of the FIB created UtS surface as a platform for circuit analysis with the main emphasis on backside electron beam probing.

2. The Focused Ion Beam Tool

This chapter will provide an introduction into the working principle of a FIB tool, followed by a general discussion of the frontside- and backside FIB CE approach, and their costs and benefits for today's semiconductor industry.

2.1. Basic FIB Configuration

As the tool name already implies, the focused ion beam tool makes use of a charged particle beam and is therefore only applicable under vacuum conditions. Figure 2-1 illustrates the necessary components of any common FIB system.

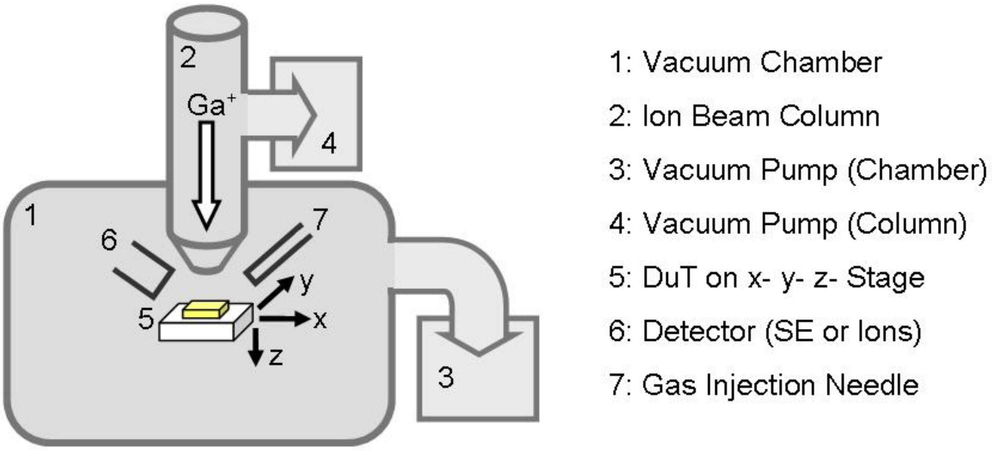

Figure 2-1: Basic FIB configuration

The framework is the vacuum chamber, usually a metal box, the size and shape of which is determined by the later field of application of the FIB system. The ion beam column is attached to the chamber. Most current systems run a three stage vacuum pump system:

- Roughing Pump: Evacuates the chamber through the Turbo Pump (down to ≈ 10^{-2} Pa)

- Turbo Pump: Evacuates the chamber (down to ≈ 10^{-5} Pa)

- Ion Getter Pump: Attached to the ion column (down to ≈ 10^{-7} Pa)

The DuT is mounted on an x-y-z-stage in close proximity to the pole-shoe of the ion column. To allow for imaging, the system has to be equipped with a detector unit, either detecting secondary electrons (SE) or secondary ions, depending on the detector bias voltage polarity. If gas assisted operations are considered, a gas delivery system has to be included, allowing the positioning of an injection needle close to the DuT surface.

2.2. Ion Beam - Sample Interaction

Most commercially available FIB CE systems use Gallium$^+$ ions and are optimized for use at an acceleration Voltage of 30 kV. Applying such a Ga ion beam to a material surface (e.g. Silicon) has several effects/interactions, illustrated in Figure 2-2.

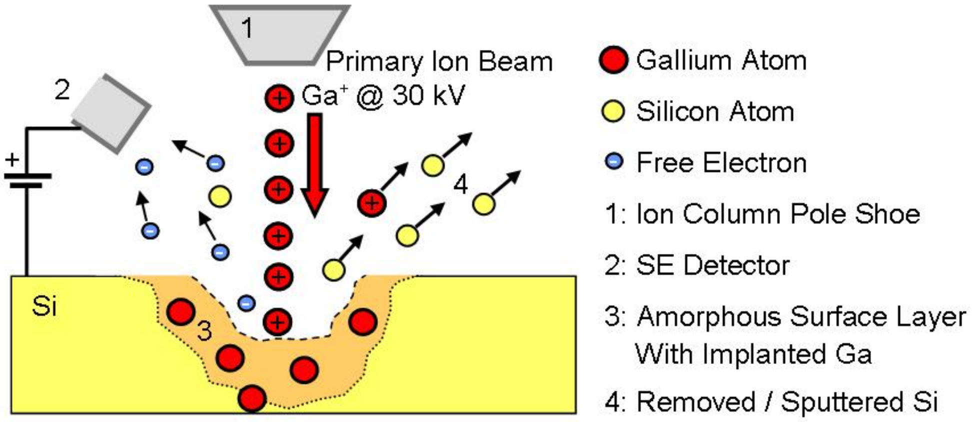

Figure 2-2: Ion beam - sample interactions

The high energetic ions bombard the surface, destroying the crystalline structure and sputtering away the material. As a result, the ion beam can be used to remove material in a direct writing manner. The remaining material will show an amorphous surface layer (≈ 30 nm, [14] & [15]) in addition to Gallium contamination. Furthermore, with every collision of an incident ion and a Si Atom, there is a high probability for SE emission. Figure 2-2 shows the classical configuration for SE detection, where the detector is positively biased compared to the sample surface, increasing the collection efficiency of a given system.

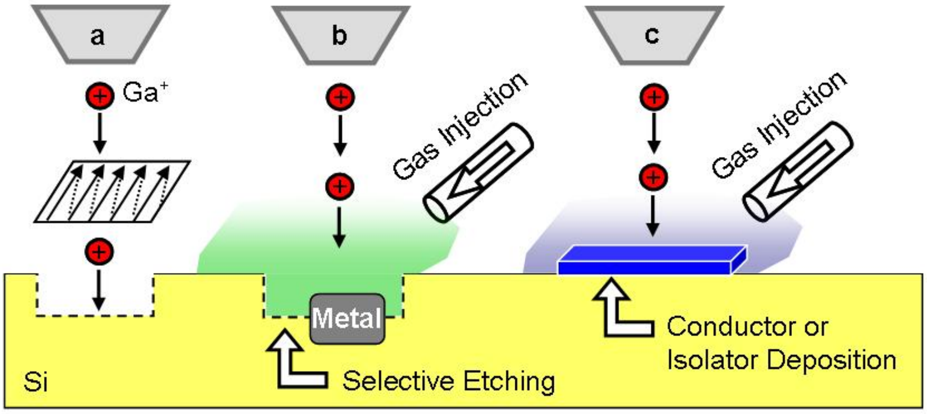

Figure 2-3: Ion beam scan pattern and principle of gas assisted FIB operations

During a FIB operation, the focused ion beam is scanned across the surface in a predefined area (box), following a scan pattern as shown in Figure 2-3 (a). Additionally, an increasing number of process gasses can be used in combination with the ion beam [4] & [13], allowing for higher etch speed combined with material selective etching (b) or the deposition of either isolating or conducting material (c).

2.3. FIB for Circuit Edit

As mentioned in the introduction, there is a strong demand for circuit modifications (or CE) in the early phase of Chip development, where repairing design failures or marginalities are often desired. Gaining functional engineering samples allows verifying design changes, delivers demonstration samples and most importantly ensures that all critical sites have been located before the final mask change.

The FIB allows modifying the interconnect layers of a given circuitry. The general process is illustrated in Figure 2-4. After locating the targeted circuit node using CAD alignment, a FIB etch operation is carried out to open an access hole to the metal interconnect (steps 1, 2 & 3). During the milling, a special etch chemistry can be used which increases the removal of the inter-metal dielectric (IMD) and reduces the etching of the metal interconnects. After that, a new electrical connection can be established by FIB induced conductor deposition. To achieve a low ohmic bridge, the operator should fill the vias first (4 & 5) before finally depositing the conductor line, connecting the two circuit nodes (6). To cut a metal line (7), a different etch gas can be utilized, selectively etching the metal and not the IMD.

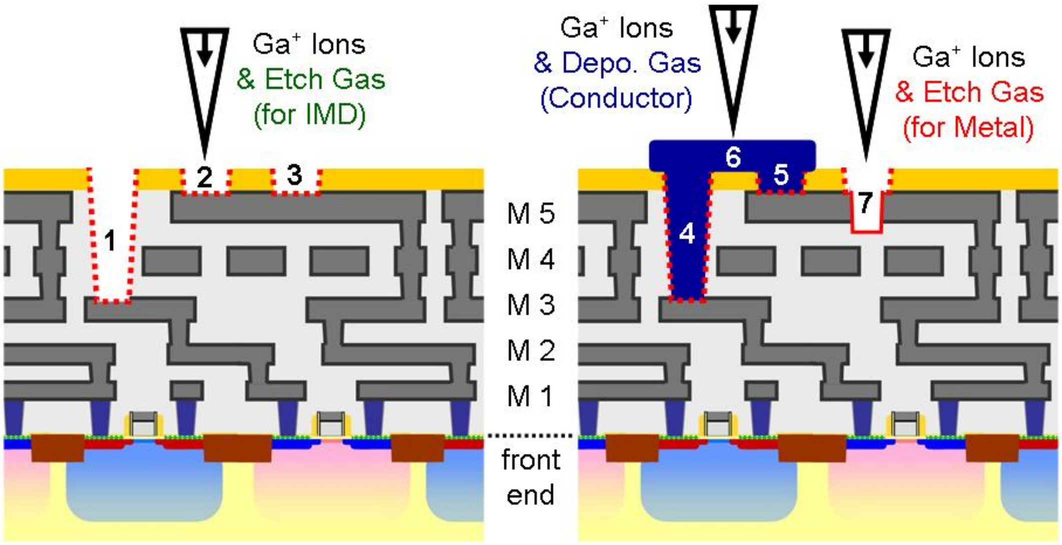

Figure 2-4: Main steps of a frontside FIB circuit edit

This process has been very successfully applied throughout the past two decades and FIB CE became a standard procedure in modern Chip development [1]. But with the ongoing technology development, the above described CE approach is facing many challenges, arising due to the ever increasing number of interconnect layers in combination with the reduction of the minimum feature size.

2.4. Challenges for frontside Circuit Edit

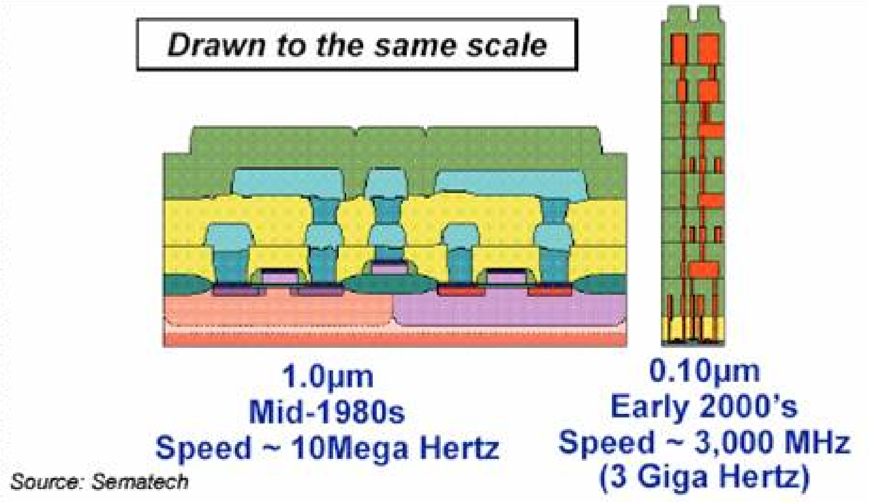

Figure 2-5: Schematic cross section through ICs of two different technology generations

The comparison between a simple 2 metal process (mid 1980s) and an 8 metal process (early 2000s) (Figure 2-5) illustrates the origin of most problems listed in the following:

- high number of metal interconnect layers
- aggressive minimum feature size, or pitch
- high aspect ratio (HAR) access holes
- fully planar process
  - lack of alignment features
  - presence of electrically floating chemical-mechanical-polishing (CMP) fill shapes
- low-k material as IMD
- potential ESD damage due to charging of the isolating passivation & IMD
- copper as interconnect metal
- package innovations (flip chip, BGA on chip…)

To take full advantage of the ongoing device (FET) shrink, the minimum size and distance (pitch) of the lowest metal interconnects have to scale down simultaneously. In addition, there is a trend in chip design to use an ever increasing number of interconnect layer to cope with chip complexity. In doing so, most of the local routing takes place on the lowest, and most aggressive (smallest) metal layers. Unfortunately, therefore most CEs also have to be done on these smallest and deeply buried structures. Designs incorporating more than three metal levels enforce planarized metal processes, where additional and not electrically used fill shapes (CMP fills) are added to the final layout to ensure a certain percentage of metal coverage in any area. As a result, trying to open a hole to one of the lower metal layers, without touching or destroying upper metal layers, often enforces HAR access holes (aspect ratio > 20 [16]) or may as well be impossible. If the access is only blocked by CMP fill shapes, the operator can mill right through them, not having to worry about any electrical disfunctionality, but the challenge is to remove these floating structures in a planar manner.

Especially with Copper (Cu) as interconnect material, this can be very time consuming and special gas chemistries are needed for reliable edits (cut operations on the bottom of HAR holes are the most critical operations regarding this concern). Furthermore, in case of a desired conductive connection to a low level metal, the HAR access holes have to be filled with conductor material. Processes have been developed achieving good via-filling [17], but due to the geometry, the resulting high resistance can still be a limiting factor.

Approaching the chip through the structured frontside, all interconnects are covered by highly isolating passivation (of various material compositions - varying from technology to technology) leading to unpredictable surface charging, which can deflect the ion beam (disturbing the alignment) or end up in ESD damage, destroying the DuT in the worst case. Also the alignment itself is challenged by the planarized technology, since the uppermost level may not reveal any structural information (being perfectly planar). This enforces doing several subsequent alignment steps, including the fabrication of FIB holes to find more localized alignment structures. In addition, milling down using a top metal layer for the local alignment also has a systematic limitation: The chip itself is build, starting with the frontend with highest interlayer alignment accuracy. But climbing upwards in the interconnect levels, the metal structures get bigger and the processes can be carried out with lower accuracy (more cost effective, using older tools). As a result, the top metal levels available for alignment may already be misaligned themselves, compared to the lowest levels (or frontend). The effects of low-k material are not fully understood yet, also because the material compositions still vary a lot, but the reliable and planar removal of these new materials poses many problems [18]. All these challenges, combined with the growing importance of new package methodologies, sometimes totally blocking frontside access, led the way to backside CE.

2.5. General Backside FIB CE Procedure

The general process can be understood by use of Figure 2-6. At first, the general backside FIB process is compared to the conventional frontside approach, with special emphasis on the challenges discussed above. A detailed description of the utilized backside FIB tool and the most important process steps is given later in chapter 3.2.

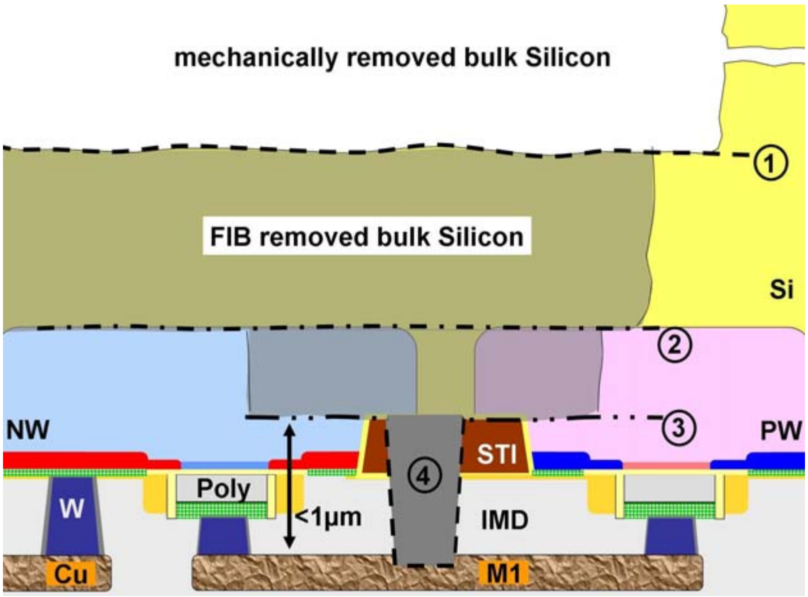

Figure 2-6: Backside FIB CE process flow

Depending on the package strategy of a given DuT, the backside preparation procedure may have to start with decapsulation to allow access to the rear (bulk) of the chip. Now the die is mechanically thinned to a remaining thickness of 10 - 60 µm, depending on the confidence level of the operator and the requirements according to the later mechanical stability and heat sink properties (1). The thinner the mechanical preparation gets, the less material has to be removed in the FIB later on. In any case, the mechanical process has to end with a high quality polish step to ensure optical transparency of the surface and to enable the following FIB procedure. Now, the chip can be transferred into the FIB vacuum chamber.

Once the DuT is in the FIB, the operator has to find the area of interest by performing a CAD alignment. After that, the FIB thinning can start, initially covering an area of up to 300 x 300 µm^2. Milling through the bulk silicon, the FIB image does not show any structural information of the underlying circuitry since the ion beam does not penetrate the matter. The first structural information is gained when hitting the n-wells at a remaining Si thickness of 2-3 µm (2). At that point, the coarse FIB thinning is stopped and a thin isolating layer can be deposited to seal the surface. The n-well contrast allows for local alignment (accuracy better than 80 nm) [10]. Now, smaller opening can be milled down to the level of the shallow trench isolation (STI) structures. The STI consists of SiO_2 and shows a very strong contrast to the bulk Si (Figure 2-7, right). This step finalized the preparation procedure and the STI shapes allow for highest local alignment accuracy [10].

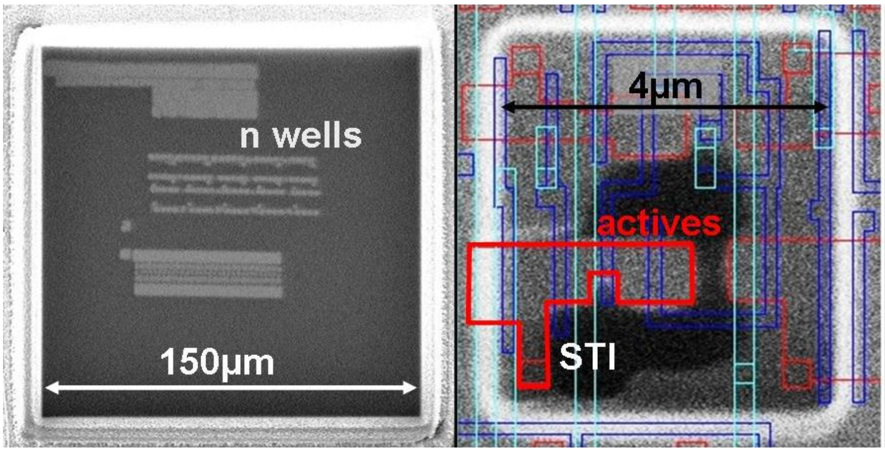

Figure 2-7: FIB images of the initial FIB trench showing the n-wells (left) and an opening to STI level including the Layout for local alignment (right)

Using the local alignment, the operator can now etch the desired access holes to the targeted metal node ((4) in Figure 2-6), milling through electrically unused areas (mostly STI), and then either cut this metal line or establish a connection by means of conductor deposition.

2.6. Comparison between frontside and backside CE

Regarding the challenges for frontside CE discussed in 2.4, the backside approach offers several advantages. The most obvious one is the better access to the lowest interconnect levels. The only blocking structures are the transistors or other metal and poly routing, but also these problems can be solved by use of new contact methodologies, later described in 9.3.

Almost similarly important is the reduced aspect ratio of the required access holes. As illustrated in Figure 2-6, the distance between the rear surface at STI level and M1 is less than 1 µm. The small distance allows for aspect ratios below 4/1 in most cases, drastically reducing the via-resistance. It also reduces the problem associated with the presence of low-k IMD and floating CMP fill shapes, since it becomes less likely to have to mill through either of those.

Also cutting Copper lines can be easier to handle either as a result of the lower aspect ratio holes or simply because the especially problematic thick power lines (top metal layers) are no longer in the way.

The local alignment accuracy using the STI shapes offers almost unlimited accuracy since the STI is clearly visible (showing a strong material contrast to Si) and has minimum offset to the lowest interconnect levels resulting from highest process constrains.

Charging problems are almost completely overcome because most of the FIB operations take place in between-, or with the conductive bulk Si surrounding the work field of the ion beam.

The biggest drawback for backside CE is the necessary mechanical preparation. Regardless of the package, thinning a die to less than 50 µm is time intensive (> 3 hours, for highly trained personal), worsens the thermal management, and also poses the risk of physical destruction to the DuT.

In general, for CE on less aggressive technology (low number of metals & bigger pitch) frontside CE can still be very successful and potentially faster, but for up to date circuit designs the frontside approach is facing its limitations, whereas backside CE clearly offers the higher future potential.

3. Backside Circuit Edit

Starting with the explanation of the main differences or additional features of the used FIB tool, the second part of this chapter will discuss the most important process steps of the proposed backside CE procedure, in more detail.

3.1. The used FIB Tool - OptiFIB II

Compared to all other available FIB CE tools, the utilized OptiFIB II system (developed by the FIB group of Schulmberger, now part of DCG Systems) has additional unique features, summarized in Figure 3-1.

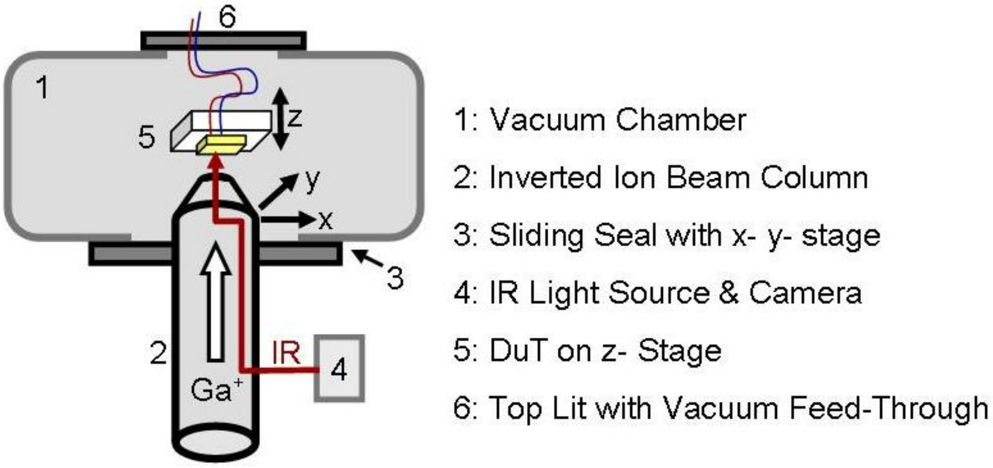
6
1
5
z
y
x
3
2
IR
4
Ga+
1: Vacuum Chamber
2: Inverted Ion Beam Column
3: Sliding Seal with x- y- stage
4: IR Light Source & Camera
5: DuT on z- Stage
6: Top Lit with Vacuum Feed-Through

Figure 3-1: Unique features of the utilized OptiFIB II

The most obvious, but almost unimportant difference is that the ion beam column (2) is inverted compared to most other system configurations (locking upwards).

The key feature of the tool is the additional IR microscope using "Schwarzchild-Casseganian" optics [19]. In this way, the ion beam and the IR microscope work in a coaxial way, simultaneously pointing at the very same spot on the DuT.

Another important difference is that the full column can move in x- & y-direction by use of a sliding seal (3), whereas the sample remains fixed, except for initially moving in z-direction to find the optical focus position. This allows to power up DuTs and to monitor their performance in-situ by use of a relatively easy setup (6).

3.2. Backside CE Process

As briefly introduced in 2.5, a backside CE process consists of the following steps:

- Mechanical Thinning

- Start of FIB work -

- Global Navigation / CAD Alignment
- Trench to n-Well Level
    - Silicon Surface Cleaning
    - Coarse Bulk Removal
      (Planarity Control)
      (Thickness Control)
    - n-well Endpoint
- Trenches to STI Level / STI Alignment
- Circuit Modification

The mechanical thinning was carried out with standard tools to a targeted remaining Si thickness of 20-50 µm and is not part of this work.

3.2.1. Global Navigation / CAD Alignment

Starting with the DuT in the FIB vacuum chamber, the die area is localized using the ion beam with minimum magnification. But since the rear side of the chip is not structured, the FIB image does not contain necessary information for a CAD alignment. To allow working with the IR-optics, the chip has to be positioned in the focus point of the optical system (DuT moving in z-direction). The chosen wavelength is 1000 nm, allowing penetrating low doped Si up to a thickness of ≈70 µm, also depending on the quality of the polished surface. Using the optical image, a first CAD alignment can be done allowing navigating to the area of interest. The FIB system is pre-aligned in a way so that the center point of the optical system coincides with the center point of the FIB image. As a result, being able to localize the area of interest optically allows for direct start of the necessary FIB work at the very same spot. The alignment accuracy of the optical system depends on the sample thickness, surface quality and also on the structures being visible on the lowest circuitry levels (poly and M1). Under average circumstances the optical alignment accuracy should be better than 1µm [10].

3.2.2. Trench to n-Well Level

Silicon Surface Cleaning

Before the major part of the bulk silicon can be removed using the most aggressive etch-chemistry, it is critical to clean and planarized the surface. The most effective way to do so is to apply a sequence of pure sputter operations (no additional process gas) and gas assisted etch operations using the less aggressive aluminum etch gas (Iodine based). The cleaning procedure is applied to the whole area of interest (up to 300 x 300 µm^2) using the maximum ion current of 20 nA (with 30 kV acceleration voltage, which is the default setting and always selected if not stated otherwise). During sputtering, the surface is turned amorphous and due to an increased sputter rate with non-normal incident angle any elevated structure or particle will be removed faster than its flat surroundings. The following gas operation removes most of the amorphous layer, increasing the optical transparency of the surface. The ratio between sputtering and gas assisted etching is ≈ 3/1. Figure 3-2 shows the progress of such a cleaning procedure where the surface quality was considered to be acceptable after 3 cycles.

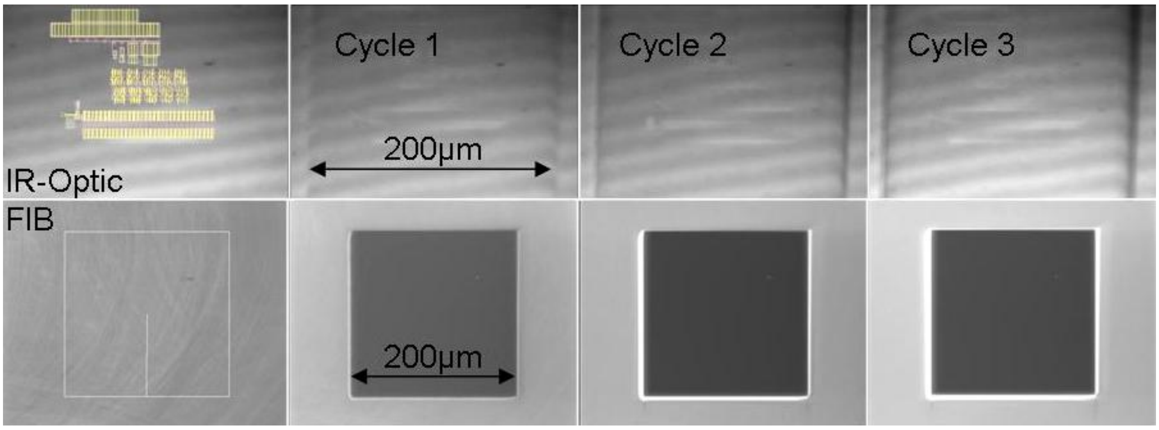

Figure 3-2: IR-Optical- and FIB images showing the 200x200 µm^2 area of interest before and after three subsequent cleaning cycles with 3 min cut and 1min Metal-etch at 20 nA beam current per cycle

Coarse Bulk Removal

Once the surface is cleaned, the rest of the bulk material has to be removed in the area of interest. For minimum process time, the most aggressive etch chemistry (Xenondifluoride, XeF_2) is used together with the highest beam current (≈ 20 nA), offering an average etch rate of 1.5 - 2 µm/minute, assuming a 200 x 200 µm^2 wide FIB box. This translates into a Si removal rate of 6 - $10\cdot10^4$ µm^3 /min, or 750 - 1000 Si-Atoms per Ga$^+$ ion.

Throughout this process step it is critical to control the remaining Si thickness and the trench planarity. A pure FIB system can not deliver the necessary information since the Ions do not interpenetrate the matter. Using the OptiFIB IR-optics allows for both, thickness- and planarity-control and both information are accessible on the fly - during the trenching process.

The optical monitoring mechanisms rely on optical interference patterns, visible in the IR-image for Si-thicknesses below 40 µm. Due to the selected wavelength of 1000 nm, and with n_{SI} being the refractive index of Si, the height difference from one interference ring to the other is:

$$\Delta z = \frac{\lambda}{2\cdot n_{Si}} \approx \frac{1000nm}{2\cdot 3.46} = 144.5nm \qquad (3\text{-}1)$$

In this way, the flatness of the trench floor can be monitored by observing the number and movement of the interference patters throughout the etching process. Figure 3-3 shows a sequence of optical images, captured approximately every two seconds. Each maximum of destructive interference is equivalent to a certain remaining Si thickness and tracking the movement of these maxima over time allows determining the local thickness minima and maxima, in total offering all information for the construction of a precise topology map. In the shown example (Figure 3-3) the marked interference pattern (red dashed line) moves upwards with time. As a result, the thickness minimum has to be at the bottom of all images and is marked by the green dashed area. Measuring from this minimum, 5.5 interference patterns can be counted towards the upper end of the image (green dashed line), equivalent to a thickness difference (increase) of ≈ 630 nm.

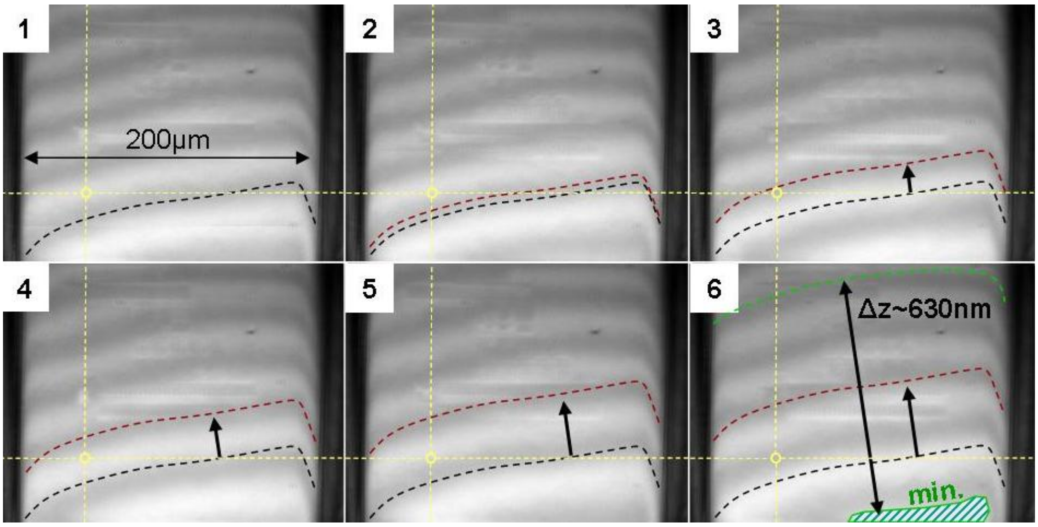

Figure 3-3: Six subsequent IR-optical images, captured during trenching with XeF_2 and 20 nA with ≈ 2 sec. time increments from one image to the other

Knowing about the unevenness of the trench floor allows carrying out corrective actions, either based on setup modifications (changing the orientation of the chip towards the gas injection needle - jet) or based on the position dependent variation of the ion beam dose. The latter makes use of the so called "Bitmap-Milling", where a 512 x 512 pixel gray-scale bitmap assigns a certain dwell time to every single pixel. Theoretically, this methodology can cope with any kind of non-planarity, but until now only linear slopes can be corrected following the standard procedure.

Monitoring the brightness of one fixed DuT location over time (e.g. the area in the center of the yellow cross-hair in Figure 3-3), the passing interference pattern create sine-waves with a relatively constant period. Assuming an etch-rate of ≈ 2 µm/min, the resulting sine-waves would have a frequency of $f_A \approx 0.24$ Hz, drawn red in the schematic diagram (Figure 3-4, the calculated frequency f_A does not scale with the x-axis). But there is a more complex side-effect of the IR-optics, superimposed to the simple interference pattern.

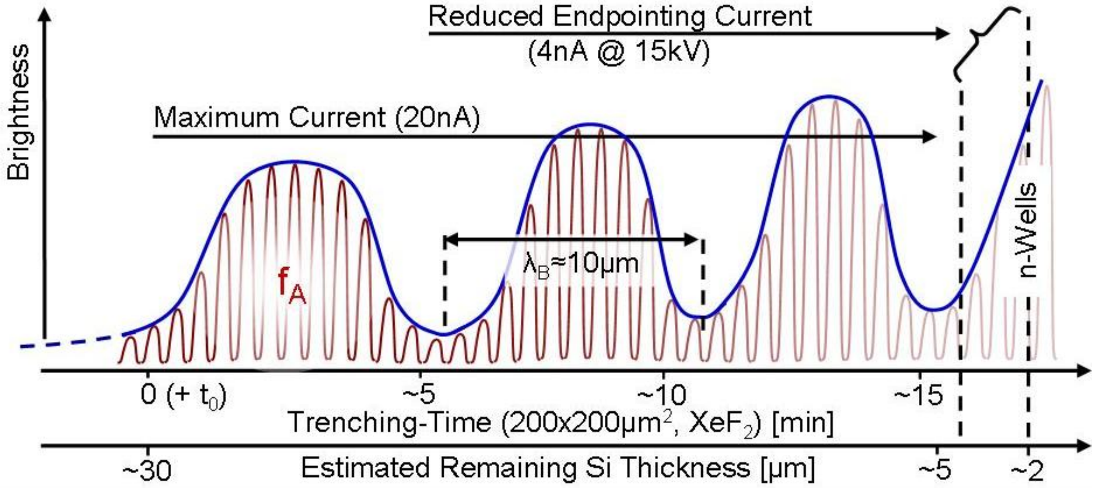

Figure 3-4: Beat progression of the interference pattern visibility as a function of FIB-trenching time and the estimated remaining Si thickness

The used IR wavelength is filtered out of a wider spectrum by means of a narrow band-pass filter with a full width at half maximum (FWHM) of approximately ±7 nm, around the center wavelength of 1000 nm, which is a relatively wide light spectrum ($\Delta\lambda \geq 14$ nm), compared to a laser-light source. This wider light spectrum initiates a beat-phenomenon (as known from acoustic waves), which is superimposed to the optical interference as a second-order effect. As a result, the visibility, or amplitude of the interference pattern (red) becomes also a function of the remaining Si thickness (the blue enveloping curve represents the superimposed beat-phenomenon).

With the used FIB system, a full period of the beat-progression is fulfilled after removing ≈ 10 µm of the bulk Si. This period can be explained by using the cut-off frequencies of the filtered IR light ($\lambda = 993$ & 1007 nm $\rightarrow$ $f \approx$ 3.021 & 2.979 THz). The resulting beat-progression would have a frequency of $f_{beat} = f_{max}\text{-}f_{min} \approx$ 42 GHz, translating into a vacuum wavelength of ≈ 71 µm. Regarding that the light passes twice through the remaining Si, which has a refractive index of $n_{Si} \approx 3.46$, delivers λ_B = 71 µm/2n_{Si} ≈ 10 µm (indicated period in Figure 3-4).

In reality, the period of the beat-progression is itself a function of the remaining Si thickness because of the changing absorption properties of the bulk with its removal. But, since the optical properties contributing to this process (mainly the refractive index of the bulk Si and the bulk to SiO_2 interface at the device level) are almost identical for any given Si chip, the characteristic beat-progression vs. Si thickness of a given OptiFIB system does not vary from chip to chip and can be used to estimate the remaining Si thickness during the trench process. This can greatly increases the confidence level throughout trenching and therefore allow speeding up the process. Using the described thickness monitoring, the coarse bulk removal can be stopped relatively late, with approximately 5 µm remaining Si thickness, without taking the risk of destroying the device layer (frontend) by accidentally milling into it.

n-Well Endpoint

The remaining material down to n-well level is then removed using a special endpointing current (4 nA @ 15 kV). The reduced acceleration voltage increases the n-well contrast and reduces the gallium implantation into the sample surface [10]. Figure 3-5 shows FIB images captured at the very end of the bulk removal (n-well endpoint, left).

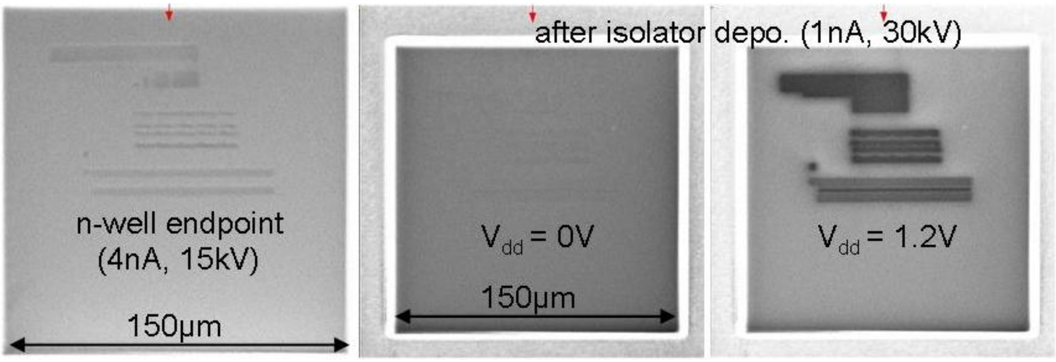

Figure 3-5: FIB images of n-well contrast at two different currents, with and without isolation and V_{dd}

Once the n-wells are visible in the FIB image, the operation is ended and a thin isolation layer can be deposited (SiO_x, using 4 nA @ 6 kV) to seal the surface, which is part of the standard procedure. If possible, it is advantageous to statically power up the device in the FIB chamber, since this greatly amplifies the n-well contrast and also ensures the best possible grounding of the bulk, preventing any charging or ESD issues. The two images in the middle and on the right show the same DuT after the isolator deposition with and without V_{dd}, demonstrating the strong increase of the well visibility caused by the supply voltage.

The n-well contrast can also be used for a local alignment with an accuracy of up to 80nm, depending on the process tolerances of the well implants and the resulting doping profiles after annealing.

3.2.3. Trenches to STI Level / STI Alignment

For highest alignment accuracy, smaller trenches can now be opened, by use of reduced beam current (≥ 10 pA/µm^2) and less aggressive Iodine based etch-chemistry, to expose the bottom of the STI shapes. The STI consists of SiO_2 and contrasts very strong with the remaining bulk Si (see Figure 3-6). These shapes offer almost unlimited alignment accuracy since they are fabricated with the smallest process tolerances guaranteeing lowest possible misalignment to lower metal levels.

Following a conservative approach, three of these boxes are opened in close proximity to the targeted edit location (STI openings only at the marked points A, B and C in Figure 3-6, the bigger opening to STI level in the center of the image would not be present for this approach), followed by a local fine alignment whereby the stage remains fixed (no physical movements) and only the ion beam is shifted via beam deflection. The gained accuracy can be ≈ 30 nm but due to the additional STI openings and the alignment procedure this approach is relatively time consuming.

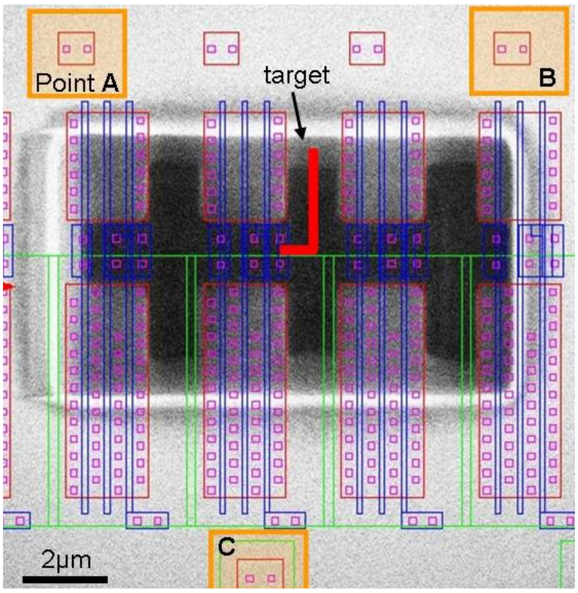

Figure 3-6: Trench to STI with CAD overlay, illustrating the two different STI alignment techniques

The faster, more accurate but more aggressive approach is also shown in Figure 3-6, where the STI is exposed directly above the targeted edit location. In doing so, the CAD overlay can be used for a local fine alignment without any further DuT movements or beam shift. Note that the revealed active areas should be covered with a thin isolating layer before any conductor deposition, to prevent short circuits.

3.2.4. Ultra thin Silicon

Only small windows to STI level are necessary for the above describe CE operations. But by use of the optical planarity control, the resulting level of flatness also allows the thinning to STI level in much bigger areas, today limited to approximately 70% of the global trench area of up to 300 x 300 μm^2. But also these limitations are not strict physical limits, more being set by a maximum process time. As discussed earlier, the remaining Si thickness of the thinned circuitry is then equal to the technology based height of the STI structures, which is close to 350 nm for all used test structures in this work. The thinned area was named ultra thin silicon (UtS) and enabled the development of novel contact strategies. Furthermore, the opportunity of creating UtS in wide areas also allows for the application of established analysis techniques like electron beam probing (EBP) and other optical techniques, directly interacting with the exposed transistor actives, offering superior resolution and adding to the variety of future FA techniques, being discussed in chapter 10.

3.3. Circuit Modifications

With the trench and the alignment being completed, the desired modification of the circuitry can be attempted. Classically, this is done by opening access holes to targeted circuit nodes on the interconnect levels. But UtS also allows to utilize different, and for many applications advantageous contact methodologies, all illustrated in Figure 3-7.

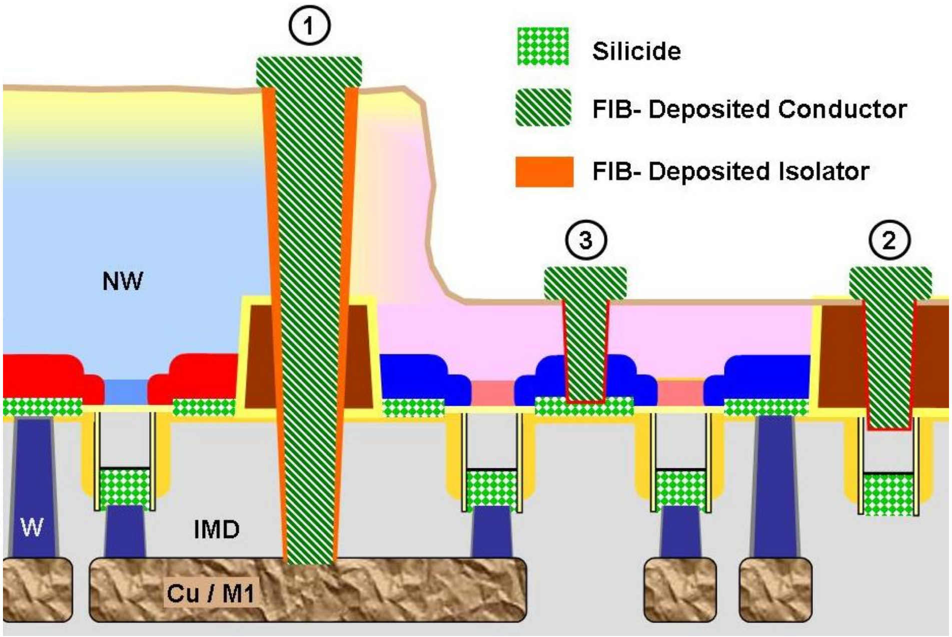

Figure 3-7: Illustration of four different methodologies to establish a contact to an inner circuit node

(1) Contact to Metal Interconnects

(2) Contact to Poly (CtP)

(3) Contact to Silicide (CtS)

For a full discussion of the advantages and disadvantages of the above depicted FIB contact methodologies, a deeper understanding of the FIB induced damage, caused by the UtS formation is helpful. Therefore, the all available options and their individual costs and benefits will be discussed in chapter 9.

4. Underlying Device Theory

This chapter will review the necessary device physics/theory regarding the MOS FET, inverter chains and ring oscillators to allow for better understanding of the later experimental results and to build the theoretical basis for the physical device simulations. If not stated otherwise, all theoretical meditations are based on the book: “Fundamentals of Modern VLSI Devices” [20], where the full derivations of the used equations can be found.

For the MOS FET, we will start with a long channel device and review the definition of the threshold voltage and the drain current for the different modes of operation. Furthermore, the substrate bias- and thermal- dependence of the threshold voltage will be briefly discussed before moving on to short channel effects, non uniform channel doping and the parasitic capacitances, leading to a brief introduction into SOI devices.

Finally, the inverter chain and the ring oscillator will be discussed in terms of their underlying working principal and parameter extraction.

4.1. The Metal Oxide Semiconductor Field- Effect Transistor (MOS FET)

4.1.1. Long Channel MOS FET

The center part of any MOS FET is the gate electrode with the gate isolator and the underlying well-material. Figure 3-1 shows the band diagram of a simplified gate structure in equilibrium (no external voltage, no current).

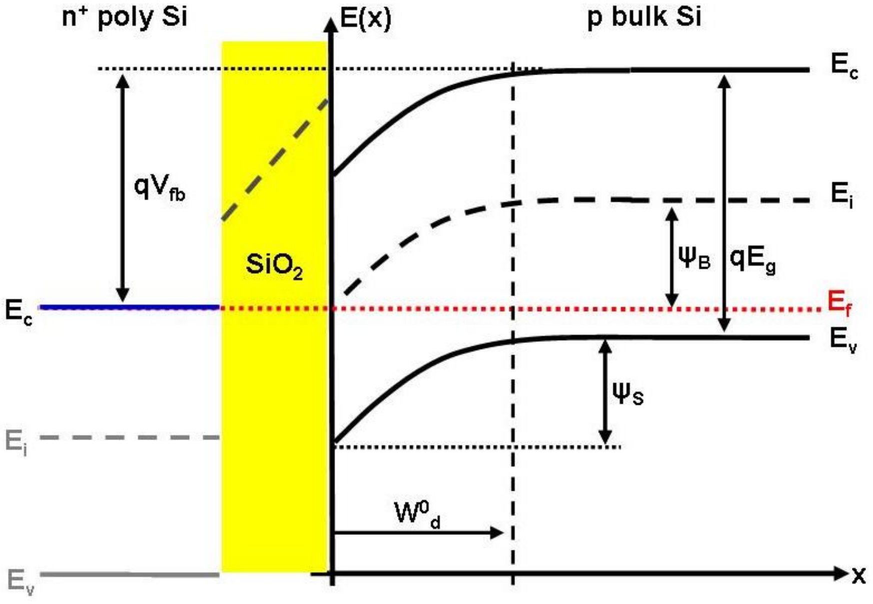

Figure 4-1: Band diagram of a simplified MOS structure with zero bias

The "metal" part of most of today's MOS FETs is comprised of highly doped poly Si, being almost as conductive as metal but allowing for self aligned processes. Poly Si has approximately the same bandgap as Si (E_g) and in case of a very high n-doping level (as present in n-type MOS FETs) the Fermi-energy (E_f) can be simplified to be identical with the conduction band energy (E_c). The p-doped well-Si has a Fermi-energy closer to the valance band (E_v), with a volume (body) potential equal to Ψ_B. If all oxide and interface charges can be neglected, the flat band voltage (V_{fb}) is equal to:

$$V_{GB} = V_{fb} = -\left(\frac{E_g}{2 \cdot q} + \Psi_B \right) \tag{4-1}$$

Applying an external voltage across the MOS structure modulates the depletion region in the well with the following dependency:

$$W^0{}_d = \sqrt{\frac{2\varepsilon_{Si} kT}{q^2 N_A} \cdot (V_{fb} - V_{GB})} \tag{4-2}$$

At the onset of strong inversion, the expansion of the depletion region reaches its maximum with:

$$W^0{}_{d\_m} = \sqrt{\frac{2\varepsilon_{Si} kT}{q^2 N_A} \cdot 2\Psi_B} \tag{4-3}$$

By use of such gate structures, complete MOS FETs can be build by adding source and drain contacts as shown in Figure 4-2.

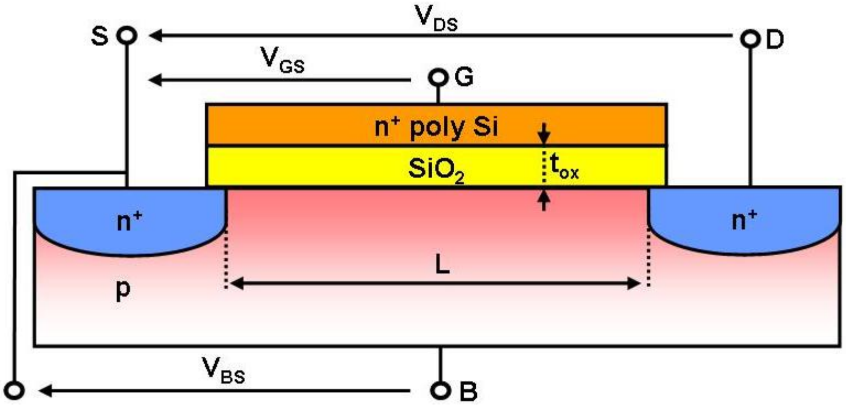

Figure 4-2: Cross-sectional view of a MOS FET structure including basic geometry parameters

With W being the width of the device, the gate to FET-body capacitance (also named oxide capacitance) C_{OX} can be derived similar to a parallel plate capacitor with:

$$C_{OX} = \frac{L \cdot W \cdot \varepsilon_{OX}}{t_{OX}} \tag{4-4}$$

The threshold voltage (V_t) of a MOS FET is defined as the gate voltage (V_{GS}) for which the band banding at the Si/SiO$_2$ interface (Ψ_S, in Figure 4-1) reaches 2 Ψ_B. For a long channel device (L > 1 µm) with V_{BS} = 0 V, V_t is defined as:

$$V_t = V_{fb} + 2\Psi_B + \frac{\sqrt{4\varepsilon_{Si} q N_A \Psi_B}}{C_{OX}},$$

or by use of (4-1):

$$V_t = -\frac{Eg}{2q} + \Psi_B + \frac{\sqrt{4\varepsilon_{Si} q N_A \Psi_B}}{C_{OX}}. \tag{4-5}$$

All definitions above include intrinsic device parameters, being possibly unknown or difficult to extract with modern short channel devices. For ease of use, several other V_t definitions evolved, deriving V_t by use of simple DC measurements. The two most commonly used methodologies are both based on a linear transfer characteristic measurement (I_D vs. V_{DS} curve, with very small drain bias, $V_{DS} \leq 50$ mV), as plotted in green in Figure 4-3. Measuring with reduced drain bias, the channel forms almost equally under the full gate and short channel effect can be neglected. The first extraction method is to set a fixed threshold current (I_0), which is scaled with the transistor width and gate length to $I_{th} = I_0 \cdot W/L$. $V_{t\,A}$ is defined as being the gate voltage for which this current level is reached $I_D(V_{GS} = V_{t\,A}) = I_{th}$. But since I_0 is usually set to very low values (e.g. 50 nA) this definition is rather susceptible to leakage current variations.

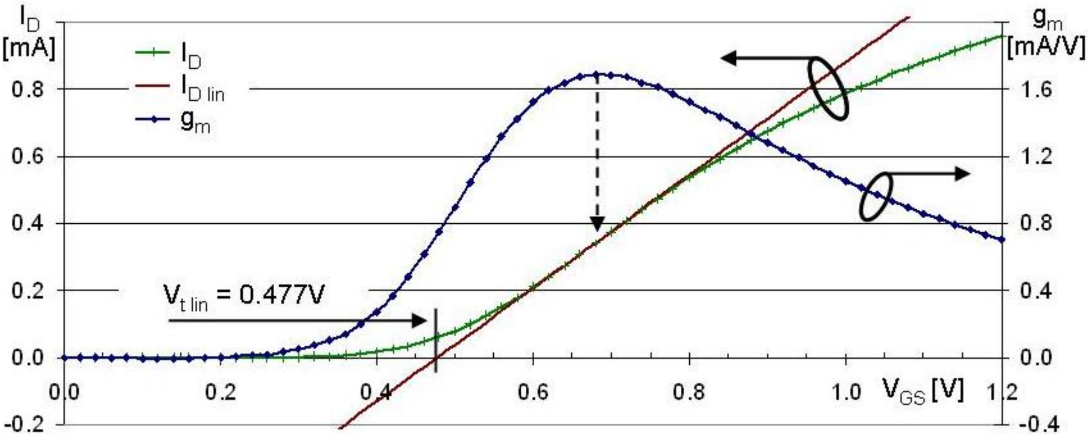

Figure 4-3: Illustration of the later used threshold voltage extraction method

The second V_t extraction scheme utilizes the linear transconductance $g_m = dI_D/dV_{GS}$. To derive V_t, a linear regression is fit to I_D at the point where g_m reaches its maxima. The so called linear threshold voltage $V_{t\,lin}$ is then equal to the V_{GS} value at the zero point of the linear regression. Due to its definition, $V_{t\,lin}$ is less susceptible to leakage currents and will be used in the later experiments. Note that $V_{t\,lin}$ is typically (2-4)·kT/q higher compared to the theoretical threshold voltage value V_t (as defined in 4-5) which is uncritical since $V_{t\,lin}$ will rather serve for comparison and process monitoring than for further calculation.

Based on the charge sheet model, the drain current I_D can be derived for the three different modes of operation

Linear ($V_{DS} << V_{GS}$-V_t) $$I_{DS} = \mu_{eff} \cdot C_{OX} \cdot \frac{W}{L}(V_{GS} - V_t) \cdot V_{DS} \tag{4-6}$$

Parabolic ($V_{DS} \leq V_{GS}$-V_t) $$I_{DS} = \mu_{eff} \cdot C_{OX} \cdot \frac{W}{L}\left[(V_{GS} - V_t) \cdot V_{DS} - \frac{m}{2} V_{DS}{}^2\right] \tag{4-7}$$

Saturation ($V_{DS} \geq V_{GS}$-V_t) $$I_{DS} = \mu_{eff} \cdot C_{OX} \cdot \frac{W}{L} \cdot \frac{(V_{GS} - V_t)^2}{2m} \tag{4-8}$$

with m, being the body effect coefficient, defined as

$$m = 1 + \frac{\sqrt{\varepsilon_{Si} q N_A / 4\Psi_B}}{C_{OX}} \ . \tag{4-9}$$

4.1.2. Definition of Gate Length

Before moving on with the theoretical meditations, it is necessary to have a closer look onto the definition of the key parameter for short channel devices - the gate length.

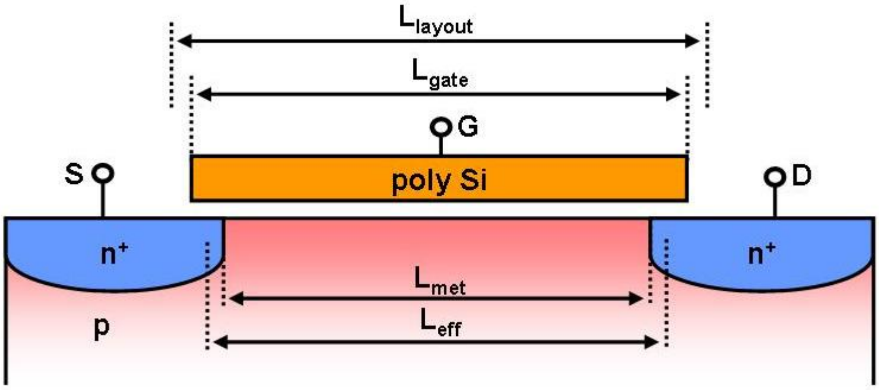

Figure 4-4: Illustration of the various definitions of the gate length

Figure 4-4 summarizes the most common definitions. L_{layout} is the assigned length of a gate structure in the layout. The resulting physical length of the poly gate L_{gate} can vary substantially from L_{layout} because of the lithography itself and following process steps like reoxidization. L_{met} is the metallurgical gate length, defined as being the length between the two pn-junctions at the Si/SiO$_2$ interface, which is almost impossible to measure for today's technology. The most commonly used gate length is the so called effective gate length L_{eff} which is different from all other definitions above since it is not a physical parameter. It is defined through the electrical characteristics of the MOS FET and is a measure of how much gate controlled current a MOS FET delivers. As a result, this definition is most suitable for circuit models and can be derived by a large number of automated measurements.

From now on, L will be equivalent to L_{eff}, if not stated otherwise.

4.1.3. Body Bias and Temperature Dependence of Threshold Voltage

Regarding a FET operating in the linear region, the depletion region below the inversion channel is reduced when the well (body) is positively biased with respect to the source potential. Assuming that the same V_{GS} still influences the same amount of charge into the Si, the reduced depletion layer width translates into an increased inversion charge density in the channel, or in other words, the threshold voltage is reduced. Hence, V_t becomes a function of the body bias V_{BS}:

$$V_t = -\frac{Eg}{2q} + \Psi_B + \frac{\sqrt{2\varepsilon_{Si} q N_A (2\Psi_B - V_{BS})}}{C_{OX}} . \tag{4-10}$$

The sensitivity of the threshold voltage regarding a substrate bias modulation is given by:

$$\frac{\partial V_t}{\partial V_{BS}} = -\frac{\sqrt{\varepsilon_{Si} q N_A \big/ 2\cdot\left(2\Psi_B - V_{BS}\right)}}{C_{OX}} \propto -\sqrt{A \big/ \left(B - 2\cdot V_{BS}\right)} \tag{4-11}$$

Differentiating V_t (4-5) with respect to the temperature delivers:

$$\frac{\partial V_t}{\partial T} = \frac{1}{2q}\cdot\frac{\partial E_g}{\partial T} + (2m-1)\cdot\frac{\partial \Psi_B}{\partial T} \tag{4-12}$$

Hence, the temperature dependence of the threshold voltage is a function of the bandgap- and volume-potential temperature dependence. The latter can be attributed to the temperature dependence of the intrinsic carrier concentration.

4.1.4. Threshold Voltage and Short Channel Effects

One commonly known model to explain short channel effects is the charge sharing model (CSM), which focuses on the depletion region underneath the channel and surrounding the source and drain diffusions (left side of Figure 4-5). With any MOS FET, a depletion region underneath the gate has to be generated by means of the gate electric field. The additional depletion regions caused by the drain and source junctions reduce the necessary gate induced depletion region at both ends of the channel.

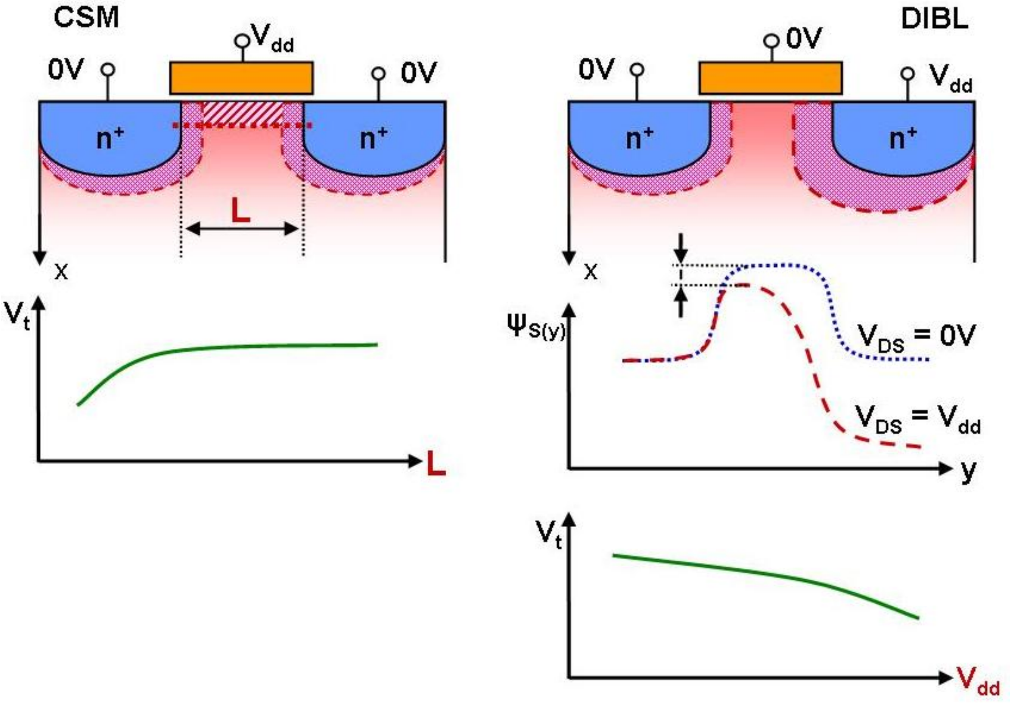

Figure 4-5: Short channel MOS FET structures including the depletion layers and left: the progression of V_t as a function of L (CSM), and right: the surface potential at the Si/SiO$_2$ interface $\Psi_{S(y)}$ and the resulting V_t over V_{dd} progression (DIBL)

For a long channel device (L ≥ 1 µm), the influence of the drain- and source junctions can be neglected, resulting in a threshold voltage as derived from the pure varactor model. For a short channel device (L <<1 µm), a big portion of the necessary depletion charge is provided by the drain/source to well depletion regions. This translates into a reduction of V_t with decreasing channel length.

A second short channel effect is the drain induced barrier lowering (DIBL), explaining the V_t dependence on the supply voltage V_{dd} with regards to the resulting surface potential at the Si/SiO_2 interface ($\Psi_{S(y)}$). At the source end, $\Psi_{S(y)}$ builds a potential barrier, limiting the injection of electrons into the channel as long as the gate voltage is below V_t. But with the short channel and the increased expanse of the drain depletion region, this barrier is strongly reduced, as illustrated in the right part of Figure 4-5, also reducing V_t.

If all doping levels are kept the same, there is a minimum gate lengths where the drain- and source well depletion regions start to overlap (referred to as "punch through"). As a result, the off current increases drastically and the FET can break down if the current is not limited elsewhere. To prevent this from happening, the doping levels have to be adjusted with the ongoing technology shrink.

4.1.5. Non Uniform Channel Doping

Regarding a uniformly doped channel of a long channel transistor, the maximum gate depletion width $W^0_{d\_m}$ (4-3), and the threshold voltage V_t (4-5) are coupled through the parameter N_A. As a result, these two parameters can not be varied separately. To control short channel effects, $W^0_{d\_m}$ should be close to, or grater than L/(2m), with m being the body coefficient derived in (4-9). But a doping level that satisfies the $W^0_{d\_m}$ criteria may not deliver the desired V_t anymore. To cope with that, non-uniform doping profiles add an additional degree of freedom to optimize device performance.

Many different channel and well-doping strategies evolved throughout the past two decades. But since all test structures used in this work are based on the same strategy, only the incorporated doping profiles will be discussed. Furthermore, the non uniform nature of most of the present implants does not allow for a complete analytical derivation of the intrinsic device physics. And since the aim of this work is not to optimize MOS FET processes, the following subsections will only highlight the general purpose of the single implants and briefly discuss their influence on the transistor performance to allow the understanding of the later used test structures and simulations.

Figure 4-6 shows the doping profiles of a simulated 120 nm FET, as will be used later for the reproduction and discussion of the experimental results. Here, this structure only serves as an example to summarize and briefly discuss all doping implants in their process given order, present in the real devices.

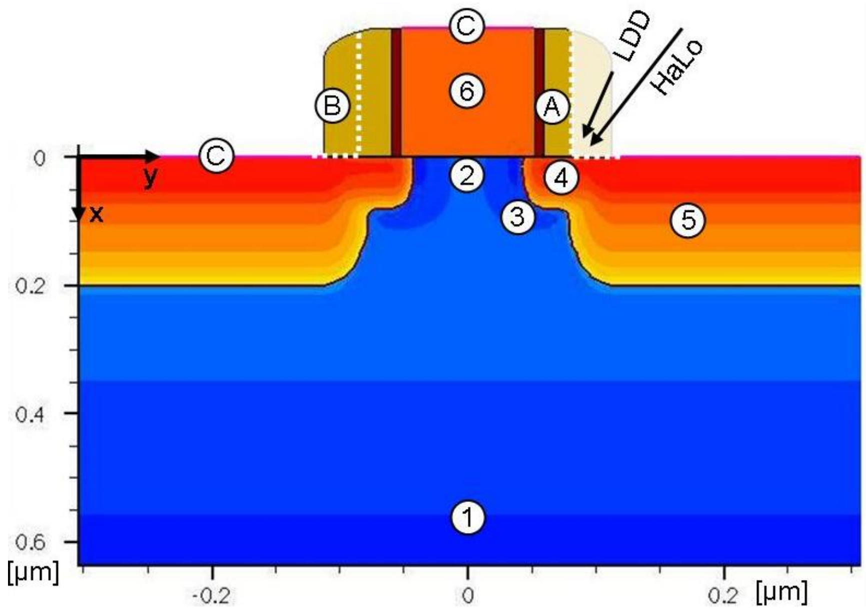

Figure 4-6: Illustration of the later used doping strategy (120 nm n-FET)

(1) Retrograde Well

After the fabrication of the shallow trench isolations, separating the later FET actives from each other, the retrograded well profile is the first doping implant. Initially, theses profiles were used to suppress parasitic bipolar devices. At that time, the well implants were buried deep underneath the channel region, not directly interfering with the intrinsic device parameters. But approaching the physical limit of the SiO_2 gate oxide thickness, t_{ox} could not be scaled to the same extent as the rest of the device anymore. As a result, there was a need to decrease V_t at a given channel doping level. Figure 4-7 shows the progression of the electric field $\xi(x)$ in the depletion region underneath a gate structure. In both cases, the situation is drawn for $V_{GS} = V_t$, where the area underneath $\xi(x)$ is equivalent to $2\Psi_B$, resulting in a surface potential of $\Psi_S = 2\Psi_B$. Compared to the uniform doping (left), the low- high- doping profile (serving as a simplified example for a retrograded well) on the right has a reduced maximum field at the surface. Since the surface field $\xi(x{=}0)$ for $\Psi_S = 2\Psi_B$ is directly proportional to the threshold voltage, this illustrates the effect of V_t reduction whereas $W^0_{d\,m}$ increases.

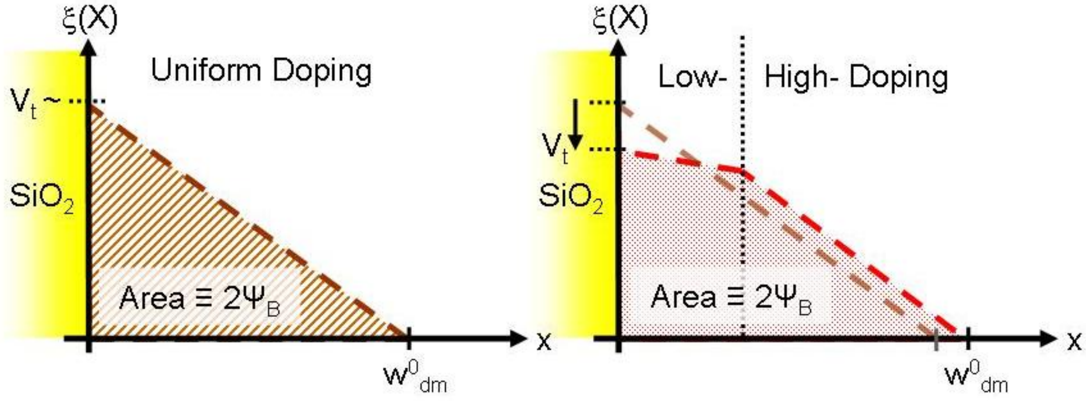

Figure 4-7: Electric field drawn over x, (left) for a uniform channel doping and (right) for a simplified retrograded well profile (low- high- doping profile)

(2) V_t Implant

Using the same mask, the V_t implant is a very shallow p-implant (for n-FETs), only effective in the later defined channel region underneath the gate. This implant is widely used to fabricate devices with different targeted V_t values by only varying this implantation dose. Most short channel effects coming with the ongoing device shrinking (e. g. CSM, and DIBL) tend to reduce V_t. In a given technology framework, varying the V_t implantation dose is one of the very few ways to bring V_t to the targeted value. But higher channel doping decreases the carrier mobility in the channel, directly decreasing the device performance. As a result, the channel doping should be kept as low as possible.

(A) Gate Oxide, Poly Gate + ReOx + Silicon-Nitride Spacer (Si_3N_4)

Following the first two implantation steps, the gate oxide and the poly gate is fabricated. Then, the sidewalls of the poly structures are oxidized (ReOx) and an Si_3N_4 spacer is formed besides the gates. The spacers play an important role in the self-aligned process, by blocking any further implantation underneath. The underlying real devices are fabricated with two subsequent spacer formations where the initial and smaller one (A) is used for the following Ldd and halo implants, whereas the final spacer (B) serves for the drain source implant.

(3) halo / Pocket Implant

In opposite to the V_t implant, the halo- or pocket implant offers a way to control the short channel effects (in other words: to increase V_t and prevent a punch through) without further degradation of the carrier mobility in the channel. The goal of this implant is to reduce the influence of the drain to well depletion region onto the channel. The implant is carried out with the same doping type as the well (p-doping for an n-FET) but on a high incident angle ($\approx 30°$), with reduced dose and moderate implantation energy (< 150 keV). The described implant increases the well- or channel doping only at the drain and source end of the channel.

(4) Lightly Doped Drain (Ldd)

The so called Ldd is an additional implant primarily used to reduce the maximum electric field at the drain end of the channel, suppressing hot carrier generation and related reliability degradation. Additionally, the series ohmic resistance from the distant drain and source to the channel is minimized. Compared to the later drain/source implant, it has a lower implantation dose and its concentration maximum is closer to the surface. The Ldd is also done self-aligned by use of the same spacer (and lithography mask) as for the halo, but with a smaller incident angle ($\approx 10°$, as illustrated in Figure 4-6)). Since this implant party overlaps with the poly gate, it reduces the effective gate length, also reducing V_t. To gain a low series resistance in combination with a well controlled V_t, the desired profile has to reach underneath the gate with a sufficiently high doping concentration and should then fall off very steeply, below the channel doping level in only a few nanometres. Compared to the halo doping, the Ldd implant has to be shallower to ensure that the halo implant surrounds the whole Ldd area, reducing the width of the source drain depletion regions in channel direction.

(B) Si_3N_4 Spacer II

(5) Source Drain Implant

The source drain implant is carried out self aligned once the full Si_3N_4 spacer is fabricated. Consequently, the implant starts in a certain lateral distance to the gate and can have highest doping levels, since it dose not interfere directly with the channel region and does not affect the intrinsic device parameters.

(6) Poly Gate Doping

Since the gate contact is fabricated out of undoped polycrystalline Si, the gate material has to be doped to achieve the desired conductivity and Fermi level. The gate is doped by use of the same implant as source and drain. To avoid depletion effects in the poly gate it is desired to gain highest possible active doping concentrations > 10^{19} cm^{-3}, all the way down to the poly/SiO_2 interface. The high doping levels also pin the Fermi level to the conduction band edge (E_c) for n-type doping or to E_v for p-type doping respectively.

(C) Self Aligned Silicide (Salicide)

The Salicide layer is also fabricated with a self aligned process and covers the top of the poly structures and every drain source area. Due to its very high conductivity it functions as a kind of equipotential plate and establishes a low ohmic contact between the Tungsten contacts and the active Si (source & drain).

4.1.6. Intrinsic & Parasitic Capacitances

The intrinsic and parasitic capacitances, summarized in Figure 4-8 are especially important for the dynamic operation of the FET in a circuitry. Out of all shown capacitances, the junction capacity of the source to well diode C_{source} is the only one not contributing to the later switching performance of the transistor, since the source never changes its potential (in regular CMOS operation). All other contributors will have to be evaluated.

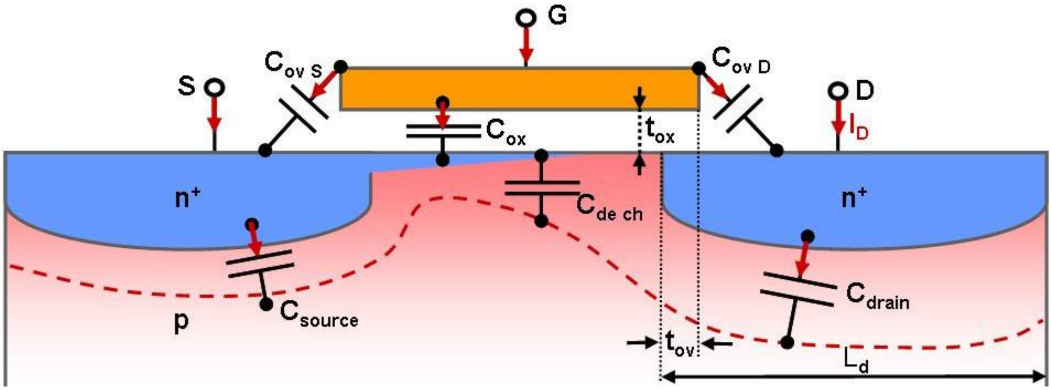

Figure 4-8: Overview of the intrinsic- and parasitic capacitances of a MOS FET, including the definition of the relevant current directions (red arrows, later used in chapter 8)

The intrinsic gate to body capacitance C_{GB} is given by a series combination of the oxide capacitance C_{ox} and the depletion capacitance $C_{de\ ch}$. But since both contributors are strongly dependent on the gate voltage, C_{GB} varies with the different regions of operation:

$$C_{GB} = WL \cdot \left(\frac{1}{C_{OX}} + \frac{1}{C_{de\_ch}} \right)^{-1} \tag{4-13}$$

Sub-threshold:

In the sub-threshold region, the inversion charge is negligible. Only the depletion charge has to be supplied when the gate voltage is changed. Therefore C_{GB} can be simplified to:

$$C_{GB} \approx WL \cdot C_{de\_ch} \tag{4-14}$$

Linear region:

With the formation of the inversion channel, the capacitive coupling between the gate and the body is strongly reduced; hence the intrinsic gate capacitance is dominated by the C_{ox}:

$$C_{GB} \approx WL \cdot C_{OX} \tag{4-15}$$

Saturation region:

When $V_{DS} = V_{Dsat} = (V_{GS}-V_t)/m$, the inversion layer charge at the drain end is reduced to zero ($Q_{i(L)} = 0$). Neglecting the channel length modulation associated with V_{DS}, $Q_{i(y)}$ can be assumed to vary parabolically along the channel

$$Q_i(y) = -C_{OX} \cdot (V_{GS} - V_t)\sqrt{1 - \frac{y}{L}} \tag{4-16}$$

and the total inversion charge can be calculated by integrating along the channel length and width to:

$$Q_{i\_sat} = -\frac{2}{3} WL \cdot C_{OX} \cdot (V_{GS} - V_t) \tag{4-17}$$

The depletion layer capacitance in saturation is very small compared to the oxide capacitance and can be associated as being part of the much bigger drain to well junction capacitance. As a result, the gate body capacitance for saturation can be simplified to:

$$C_{GB} \approx \frac{2}{3} \cdot WL \cdot C_{OX} \tag{4-18}$$

Regarding the later quantitative evaluation of the experimental results, the drain to well junction capacitance C_{drain} will be the most important. Due to design rules according the minimum lateral distance between tungsten contacts and the gate stack, regular devices are always fabricated with a much wider diffusion area L_d compared to the gate length L (for short channel devices the diffusion length $L_d \geq 3 \cdot L$). As a result, C_{drain} is dominated by the contribution of the drain diffusions, apart from the channel region where all doping profiles are laterally constant. This allows to estimate the full junction capacitance by calculating the capacitance per unit area of the drain diffusion and finally multiply by the complete diffusion area ($W \cdot L_d$), not separately calculating the capacitances originating from the diffusion endings and Ldd. Furthermore, since the doping levels of the drain and well are very asymmetric in the vicinity of the junction ($N_{Drain} = N_D \approx 10^{20}\,cm^{-3} >> N_{Well} = N_A \approx 10^{17}\,cm^{-3}$), the calculation of the capacitance can be simplified by neglecting the depletion region on the drain side to:

$$C_{drain} \approx W \cdot L_d \cdot \sqrt{\frac{\varepsilon_{Si} q N_A}{2(V_{bi} - V_{rev.})}} \tag{4-19}$$

Hereby, V_{bi} is the build-in voltage of the junction (typically around 0.9 V) and $V_{rev.}$ represents the reverse biased voltage across the junction - in this case V_{DB}.

The extrinsic parasitic capacitances are summarized in the overlap capacitance C_{ov}. Compared to all other contributors, the influence of C_{ov} is rather small and not influenced by the FIB procedure. Consequently, C_{ov} will not be necessary for the later discussion of the experimental- and simulation results.

4.2. CMOS Inverter

The circuit diagram of a single CMOS inverter is shown on the left of Figure 4-9. Assuming that the input signal has been sufficiently long on logic low ($V_{in} = V_{ss} = 0$ V), both the n-FET and the p-FET have reached steady state conditions. Due to the low input, the p-FET channel is opened whilst the n-FET is switched off. Applying a step function to the input ($V_{in}(t = 0^+) = V_{dd}$) turns off the p-FET and the n-FET is switched on. Now, the n-FET has to discharge the common node, finally pulling the output voltage V_{out} down to zero. The down propagation delay τ_n can be measured at the point where $V_{out}(t) = V_{dd}/2$, which characterizes the n-FET dominated switching performance from output high to low. The equivalent can be done for the p-FET.

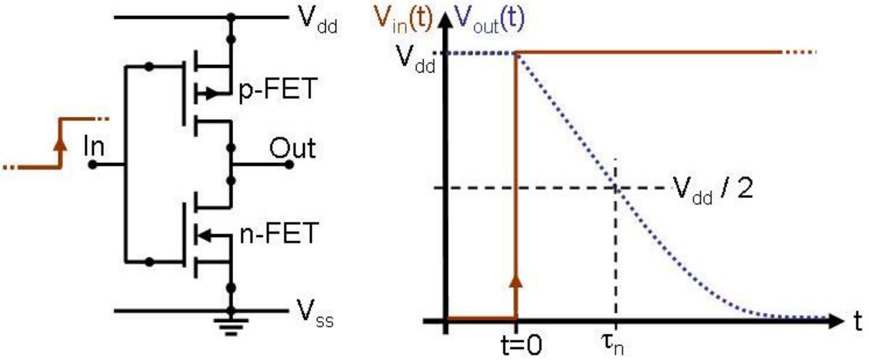

Figure 4-9: Circuit diagram of a CMOS inverter (left) and its output voltage response applying a step function at the input

But these measurements do not reflect the working conditions of CMOS gates in a real circuitry, since every gate is driven by a previous one whose output signal has a finite rise and fall time and does not deliver a step function as an input. Inverter chains are the simplest model to simulate logic operation with more realistic signal propagation, delivering a better measure of the delay times than single elements can do.

4.2.1. CMOS Inverter Chain / Switching Performance

An inverter chain is comprised of several identical inverters connected to a chain. Figure 4-10 shows the schematic of such a chain with additional load capacitors C_L attached to every single inverter output. The propagation delay is evaluated by introducing a square wave at the input. After a few stages, the signal waveform has become a *standardized signal*, i.e., one that has stabilized and remains the same shape independent of the number of stages of propagation.

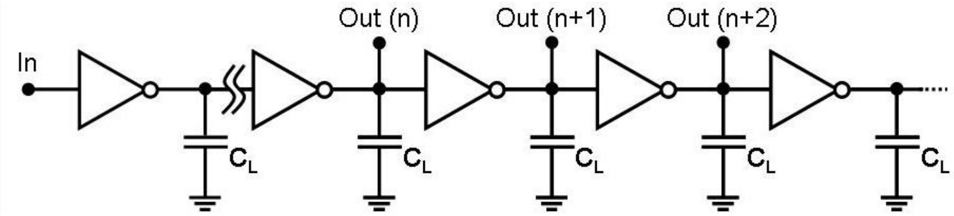

Figure 4-10: Schematic of an inverter chain with additional load capacitors attached to every output

Assuming that the signal has stabilized until stage number n, the pull up- and pull down-propagation delay can be measured, evaluating three subsequent output signals, as shown in Figure 4-11. Based on theses definitions, the average stage delay τ can be measured as defined in (4-20).

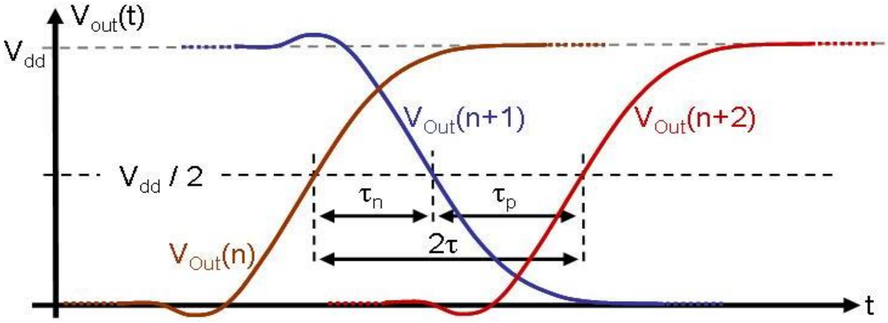

Figure 4-11: Propagation of three subsequent output signals of an inverter chain, drawn over time

$$\tau = \frac{\tau_n + \tau_p}{2} \tag{4-20}$$

Due to the complex current expressions for short channel MOS FETs and the contributing voltage dependent capacitances, numerical calculations are needed to solve the full propagation delay equations. But a simplified model based on the so called switching resistance R_{sw} allows to link all important device parameters quantitatively to the resulting CMOS propagation delay τ.

$$\tau = R_{sw} \cdot (C_{out} + FO \cdot C_{in} + C_L) \tag{4-21}$$

Here, C_L is the extrinsic (and optional) load capacitance, C_{out} is the output capacitance, summarizing all capacitive contributions connected to the common output node and C_{in} is the sum of all input capacitances of the following stage, as defined in subsection 4.1.6.

$$C_{in} = C_{GB\_n} + 2 \cdot C_{ov\_n} + C_{GB\_p} + 2 \cdot C_{ov\_p} \tag{4-22}$$

$$C_{out} = C_{drain\_n} + C_{ov\_n} + C_{drain\_p} + C_{ov\_p} + C_L \tag{4-23}$$

Note that for the input capacitance both, the overlap capacitance to the drain and to the source have to be taken into account, whereas only the drain part contributes to C_{out}. FO is the fan out, which equals one for this example.

The switching resistance R_{sw} has to be found experimentally. The most common way to do so is to fabricate various chains with identical inverters, only varying the load capacitance C_L. Plotting the resulting τ vs. C_L should show a linear dependency and R_{sw} can be determined by calculating the slope.

The main benefit of this approach is the achieved decoupling of the two important factors that dominate the CMOS performance: current driving capability (included in R_{sw}) and capacitance. On the other hand, the influence of the different n-FET and p-FET current driving capabilities can not be evaluated separately in this way. A more qualitative way to decomposed R_{sw} into R_{sw_n} and R_{sw_p}, in terms of the pull down and pull up delays τ_n and τ_p as defined in Figure 4-11 is discussed in [20] and delivers:

$$R_{sw} = \left(R_{sw\_n} + R_{sw\_p}\right)/2 \tag{4-24}$$

$$R_{sw\_n} \equiv \frac{\partial \tau_n}{\partial C_L} = k_n \frac{V_{dd}}{W_n I_n} \tag{4-25}$$

$$R_{sw\_p} \equiv \frac{\partial \tau_p}{\partial C_L} = k_p \frac{V_{dd}}{W_p I_p} \tag{4-26}$$

By use of the fitting parameters k_n and k_p, the individual contributions of the n- and p-FET can be linked to the inverse of their maximum current ($I_n = I_{D\_n}(V_{GS} = V_{DS} = V_{dd})$ and to the applied supply voltage V_{dd}. Note that k_n and k_p are also a function of V_{dd}, limiting the area of validity for the above equations (4-25 & 26).

4.2.2. Influence of CMOS Design

With every given technology there is a long set of design rules, limiting the freedom of choice for the circuit designer. Based on these rules there are a few guidelines to follow for a speed optimized CMOS design which will be briefly discussed by use of the two designs depicted in Figure 4-12.

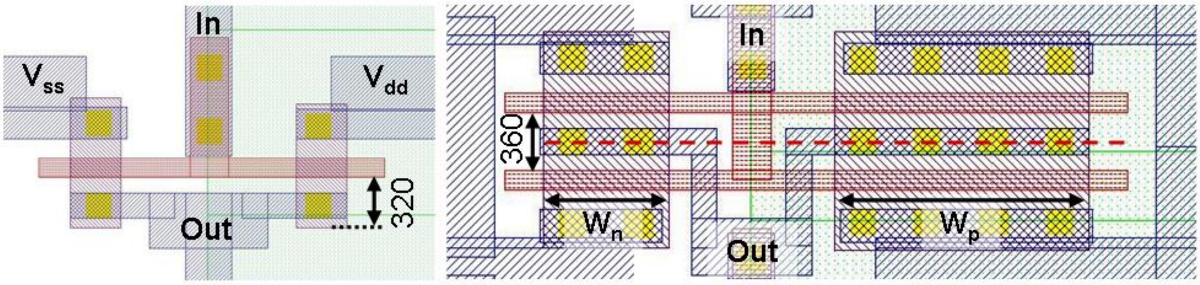

Figure 4-12: Comparison between two inverters of the same 120 nm technology with minimum active area (left) and with speed optimized (folded) design and w_n/w_p ratio (right)

The left part shows a design with minimum active area size for n- and p-FET. The source size is uncritical for CMOS performance since it is always pinned to either V_{dd} or V_{ss}. But the size of the drain is directly proportional to the drain to well junction capacitance C_{drain} (4-19), which dominates the CMOS output capacitance and consequently has a strong impact on the CMOS propagation delay τ (4-21). The width of the n- and p-FET is chosen to adjust the current driving capability of the inverter and will be discussed later. But the drain length has no positive influence and should be reduced as much as possible. For the simple design (left), the minimum drain length is given by the minimum distance between contact and gate, plus the contact size itself and the minimum extension beyond the contact, here adding up to 320 nm. The right part of Figure 4-12 shows a so called folded design where the inverter is comprised of two parallel n- and p-FETs, sharing the output (drain) contact. As marked in the image, this allows reducing the spacing between the two gates to a minimum of 360 nm.

Comparing this to the first design, the drain length associated with one transistor (red dashed line) could be reduced to 360 nm/2 = 180 nm, being approximately 56% of the initial value.

In addition to the drain length, the ratio between the widths of the two FETs is also important. The maximum current per unit gate length of a p-FET is usually only about 50% of the n-FET value. But since the switching resistance is inversely proportional to the maximum current, a p-FET, having the same size as an n-FET, would have approximately twice the switching resistance of the n-FET, resulting in an increased propagation delay of the full inverter. Depending on the individual technology parameters and the load attached to the inverter, a minimum of τ can be found for $2.5 > (w_p/w_n) > 1.5$.

4.3. Ring Oscillator

The general idea of a ring oscillator (RO) is depicted in Figure 4-13. It consists of a large odd number of inverters (here only 13) and once it is powered up a transition from high to low on one inverter enforces the transition from low to high on the next, and so on. As a result, a signal propagates around the ring at the intrinsic speed limit as defined by the technology and inverter design.

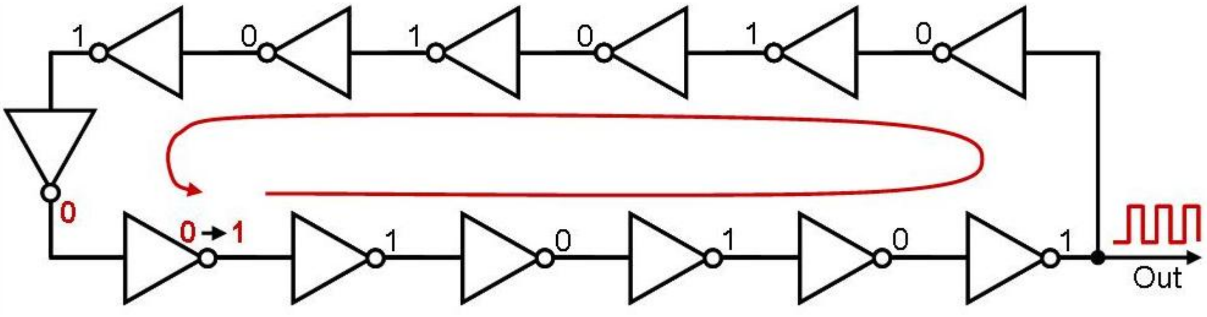

Figure 4-13: Schematic of a ring oscillator with 13 inverter stages

To allow for easy measurement of the propagating signal there must be one extra inverter leaving the ring. This gate should have the minimum size to minimize the disturbance due to the additional load it poses to the ring. Detecting the signal at this output, the transition in the ring has to propagate twice around the whole ring before a whole signal period is fulfilled. Consequently, the period of oscillation measured externally is given by

$$\tau_{osc} = 2 \cdot n \cdot \tau \tag{4-27}$$

where n is the number of stages and τ is the stage delay as defined in (4-20 & 21).

In most cases, the internal signal is coupled out, transformed into a lower frequency by use of multiple frequency dividers and finally amplified to allow the direct use of an oscilloscope.

4.4. Silicon on Isolator FET

As discussed above, the speed of a switching process is directly proportional to the current driving capability of the FETs and the capacitive load attached to the output node. Due to (4-21), the sum of all contributing capacities was found to be:

$$\Sigma C = \left(C_{out} + FO \cdot C_{in} + C_L\right) \tag{4-28}$$

For the unloaded case ($FO = 1$, $C_L = 0$) the intrinsic capacitance of the gate (C_{GB}) and drain (C_{drain}) dominate this equation. Figure 4-14 shows a comparison between a regular bulk device, a partially depleted (PD) and a fully depleted (FD) SOI FET. For both SOI structures, the buried oxide (BOX) separates the transistor body from the bulk and also prevents the formation of depletion regions underneath source and drain. As a result, the junction capacitances do not contribute any more.

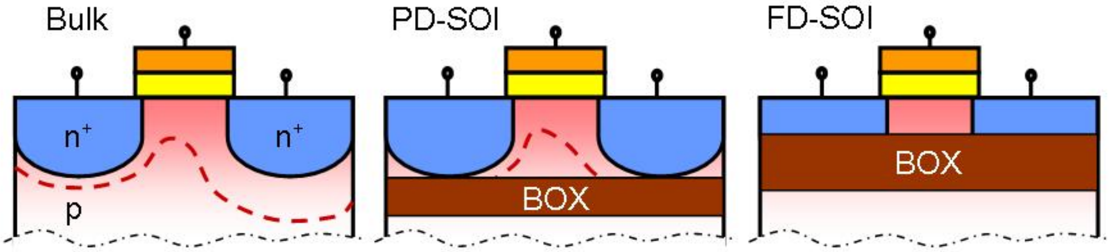

Figure 4-14: Cross-sectional view of a bulk MOS FET compared to a partially depleted (PD) silicon on insulator (SOI) FET and a full depleted (FD) SOI FET. The dashed red line indicates the maximum width of the depletion region

Eliminating C_{drain} results in a speed gain of > 20% for the unloaded case. For FD-SOI FETs, the remaining active material is further reduced to ensure that the remaining body is fully depleted in any region of operation. Consequently, also the gate depletion capacitance can be neglected.

For the interpretation of the later experiments, some effects associated with PD-SOI FETs will be important, hence they have to be discussed in more detail. As can be seen from the sketch, compared to FD-SOI, the PD-SOI active layer is thicker and the depletion width in the remaining body does not reach the BOX and is modulated with the gate potential. Since the body is not grounded, this part is often referred to as floating body, having the floating body potential φ_b.

At first the SOI FET is only considered in static operation. Similar to bulk FETs, the electric field at the drain end of the channel region reaches very high values ($> 1 \cdot 10^5$ V/cm) when the channel is pinched off in FET saturation. Now, the channel carriers (electrons for n-FETs) can gain sufficient energy to cause impact ionization (ii) when scattering. The generated electron hole pairs are separated in the high field and the electrons increase the transistor on-current. With a bulk FET the additional holes would drive a small bulk current, dissipating via the well-contact but with the SOI FETs the body is floating and the additional holes accumulate and increase the body potential. Such an SOI structure driven in static on-state is depicted on the left of Figure 4-15. The center of the image shows the resulting φ_b as a function of V_{DS} for different values of V_{GS}.

For low drain source voltages (A), φ_b is dominated by thermal generation of carriers in the drain to well depletion region and remains on a very low level. Further increasing V_{DS} drives the FET in saturation initiating ii in the pinched off channel, delivering a much bigger current and therefore also a stronger impact onto φ_b (B).With lower V_{GS} the pinch-off condition is reached at lower V_{DS} values, explaining the earlier onset with smaller V_{GS}.

As already discussed for the long channel bulk FET, any increase of the body potential decreases the threshold voltage (4-10), leading to an increase of the drain current. This phenomenon was named *Kink Effect* [21] and is best visible in the output characteristic of an SOI FET as depicted in the right part of Figure 4-15.

But substantially increasing φ_b drives the well source diode in forward direction which initiates a diode current based on the injection of electrons into the floating body, recombining with the additional holes. This limits φ_b to values clearly below the build-in voltage of the junction V_{bi}.

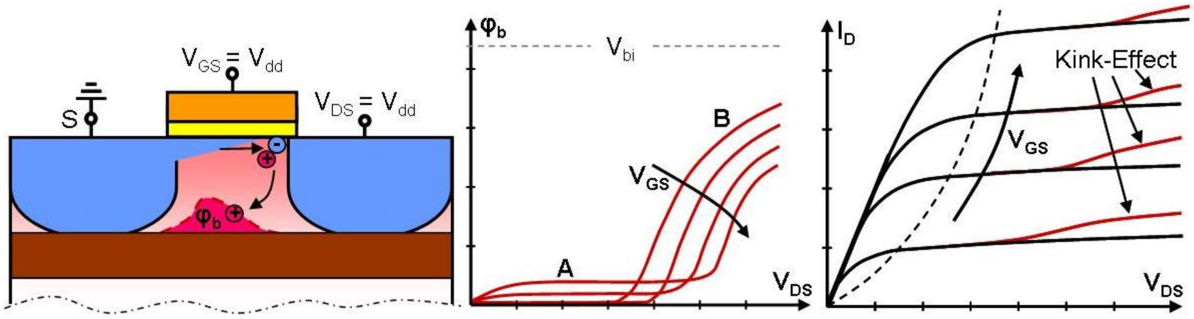

Figure 4-15: PD-SOI FET in static on-state (left), the progression of φ_b as a function of V_{DS} for different V_{GS} and the corresponding output characteristic showing the Kink Effect

In static operation, the Kink Effect only increases the maximum current and does not pose a problem. But the dynamic operation of similar FETs in a CMOS environment demands some further attention.

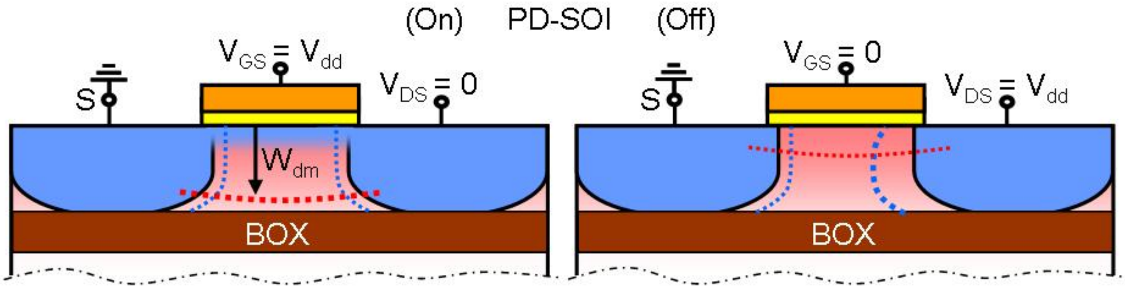

Figure 4-16: PD-SOI n-FETs including the depletion layer width of the gate depletion region (red dotted line) and the source drain junction depletion region (blue dotted line) as resulting from the applied voltages

Figure 4-16 shows two PD-SOI n-FETs with the applied voltages being equivalent to the situation in a CMOS inverter with the input statically on logic high (left) or low (right). As can be seen, either the gate depletion region is extended to its maximum W_{dm} when $V_{GS} = V_{dd}$ whereby $V_{DS} = 0$ or the drain junction depletion region is widened due to $V_{DS} = V_{dd}$ whereby $V_{GS} = 0$. Considering the input being statically on high potential (no dynamic operation), the floating body potential φ_b follows the source potential, being set to 0 in this example. φ_b may marginally increased due to the leakage current of the drain body junction.

With the event of switching, the charge conditions in the floating body change and positive carriers either have to be thermally generated in case their concentration fell below neutral, or additional carriers have to dissipate via the source well diode, driving a forward current. But both processes are rather slow compared to the operation frequency of a modern VLSI circuit and as a result, the floating body potential does not reach its static value anymore and becomes a strong function of time [22].

In general, there are four main contributors to the charge balance in the body, three of them being summarized in Figure 4-17. At first the modulation of the depletion regions: as already illustrated in Figure 4-16, the gate depletion region in the on-state fills almost the full body.

Switching the n-FET into off-state strongly decreases the gate depletion region whereas the drain depletion region increases. In the depicted example (not in general), the contribution of the gate is bigger compared to the drain leading to an increase in φ_b with the transition from input low to high and vice versa (1). Note that the increase at the positive transition has to have the same height as the later decrease at the negative transition due to the unchanged ratio between the depletion regions.

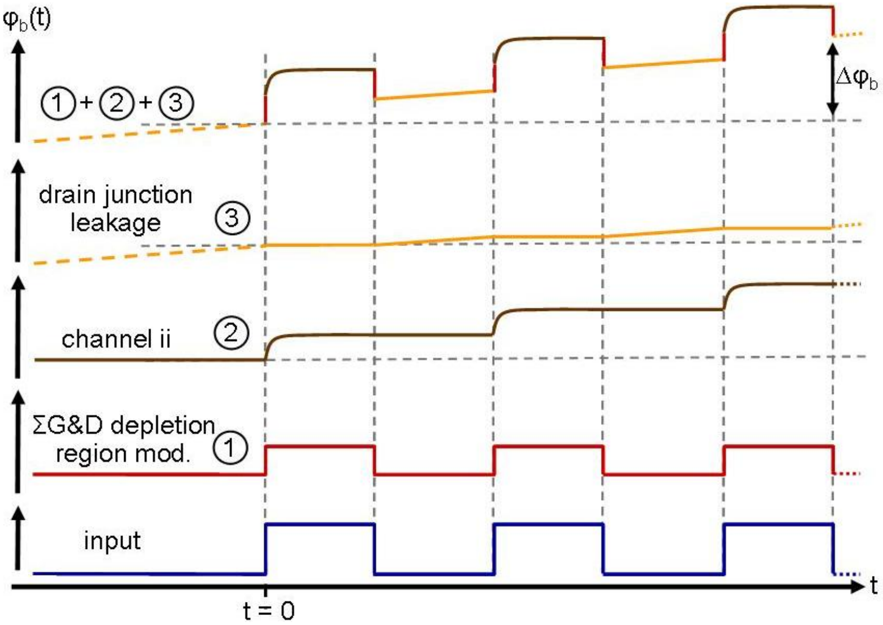

Figure 4-17: Progression of three contributions to the floating body potential and their sum, over time

The second contribution is the channel impact ionization current only occurring when the FET is driven in saturation. In a CMOS, the n-FET is operated in saturation only for a very short time, immediately following the transition from input low to high when the n-FET discharges the common output node. The channel ii (2) generates additional electron hole pairs whereby the electrons add up to the drain source current and the holes accumulate in the floating body, increasing its potential with every transition from input low to high.

The third contributor is the drain body junction leakage which is only active under input low condition, when the drain body diode has to block V_{dd}. For well designed source drain implants this part should be rather small compared to the channel ii.

For $\varphi_b(t)$, all three parts sum up and would cause a continuous increase with the ongoing switching cycles (uppermost plot in Figure 4-17). But similar to the static operation in saturation, the forward current across the well source diode limits the average floating body potential $\overline{\varphi}_b$. The system will find a balance when the amount of charge being added during one cycle is equal to the charge being neutralized by the forward diode current during the same period of time. Therefore, $\overline{\varphi}_b$ is not only a function of all previously discussed technology parameters but also a function of the operation frequency, since a higher frequency will increase the charge added per time due to channel ii. Even though the floating body effects tend to enhance circuit performance, the resulting threshold voltage reduction and drain current enhancement is history dependent. The floating body potential depends on how recently and how often a device has been switched.

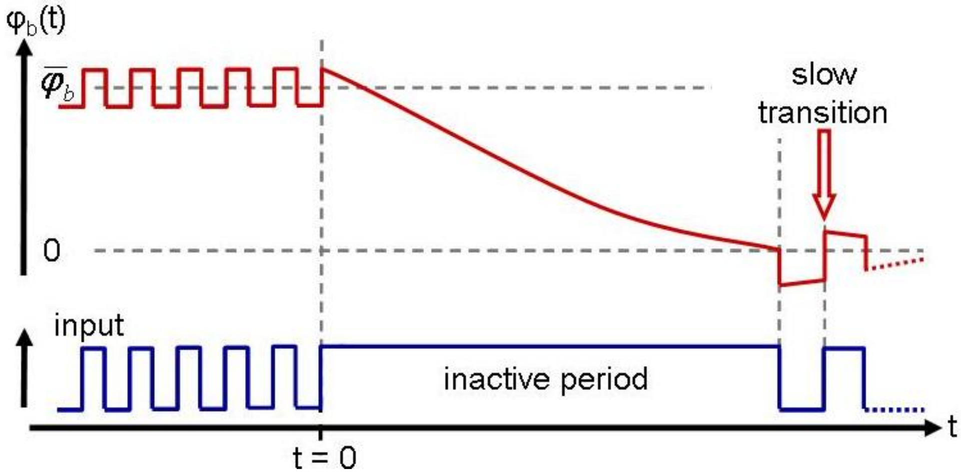

Figure 4-18: Illustration of the worst case scenario, according to the $\varphi_b(t)$ modulated switching performance of an PD-SOI n-FET, operating in a CMOS inverter

Unfortunately, performance degradation can also occur when a full switching cycle follows a long period of inactivity (as depicted in Figure 4-18, again focusing on the n-FET). Due to the long idle time, the floating body potential can drop to its static value close to zero, with the n-FET being switched on. With the transition from input high to low, $\varphi_b(t)$ drops below zero because of the depletion region modulation. This results in an increased threshold voltage, slowing down the next transition from input low to high. After a short period of continuous switching, $\varphi_b(t)$ will reach the average value $\overline{\varphi}_b$ again.

Consequently, the designer has to include an additional performance margin to prevent the history effect from posing a problem to the design stability. As a result, not the full performance increase associated with the SOI technology can be utilized.

5. Physical Device Simulation

As can be concluded from the previous chapter, the physics of modern devices are by far too complex to allow for a closed analytical description. To overcome this problem, physical simulations in the Synopsys® Sentaurus environment are utilized in order to link the most important alterations of device performance to certain intrinsic device parameters. Synopsys® Sentaurus simulates numerically the electrical behavior of a single semiconductor device or several physical devices combined in a circuit. Terminal currents, voltages, and charges are computed based on a set of physical device equations that describe the carrier distribution and conduction mechanisms. This chapter will at first provide a brief introduction into the working principal of the simulation environment. Following, the derived DC modes for the later experimentally modified single n- and p-FETs will be introduced and discussed with regards to their physical relevance. Finally, the later used dynamic simulations will be briefly discussed, only serving as a basic introduction before the necessary dynamic simulation models will be derived and used in chapter 8.

5.1. The utilized Simulation Tool

Synopsys® Sentaurus is an up to date, multipurpose simulation environment based on numerically solving a given set of device equations for predefined structural elements. A full description of the tool capability and underlying physical models can be found in the manuals [23] & [24]. For the simulations in this work, only two program parts were utilized and will be discussed in the following.

5.1.1. Structure Editor

For all later simulations, the physical structures are generated by the structure editor [23]. All simulations are based on 2-dimentional models, including the dimensions and materials of all device parts (e.g. gate oxide, SiN space, poly gate, well, STI…), the outer contacts, the doping profiles and the mesh, determining the calculation points for the numerical solver.

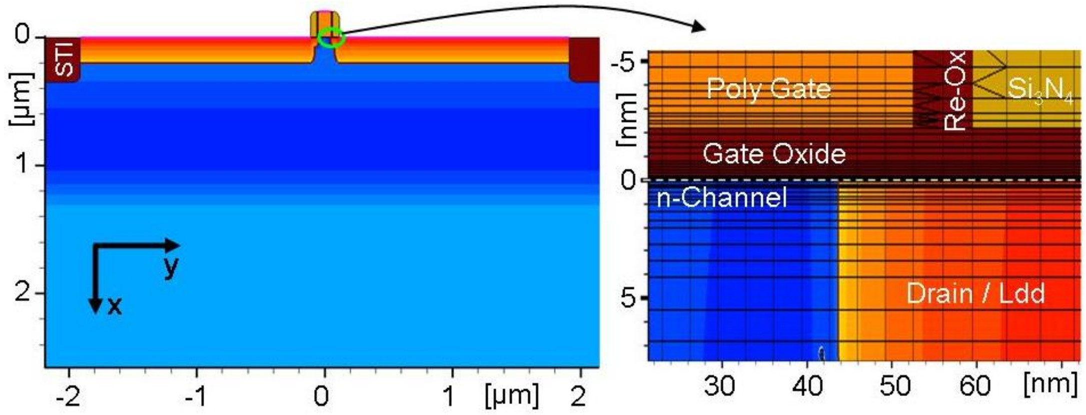

Figure 5-1: Structure of a single n-FET (left) with enlarged view of the right end of the channel area, including the mesh (right)

The generation of a test structure as shown in Figure 5-1 follows a certain order:

(1) Device Geometry and Material

All material blocks are created by defining every vertex of a rectangular and setting the designated material. If necessary, for example with the STI, a corner can later be rounded to create more realistic geometries.

(2) Contacts Declaration

A contact is defined in between two vertex of the same material block. For example, the source contact is defined on top of the left half of the test structure, in between the upper right vertex of the STI and the lower left corner of the Si_3N_4 spacer. All contact property will be defined later in the Sentaurus Device section.

(3) Doping Profiles

In combination with the device geometry, the doping profiles dominate the performance of the generated device and have to be discussed in more detail. The structure editor allows the implementation of various different profile patterns out of which only the homogeneous- and the Gaussian profiles are utilized. If more than one pattern is assigned to an area, the resulting doping is always the sum of all contributions.

With the homogenous profile, a constant doping level is assigned to a full material block. In the example FET in Figure 5-1, one homogenous profile is applied to the poly gate material and a second to the full bulk Silicon, equivalent to the minimum doping of the wafer.

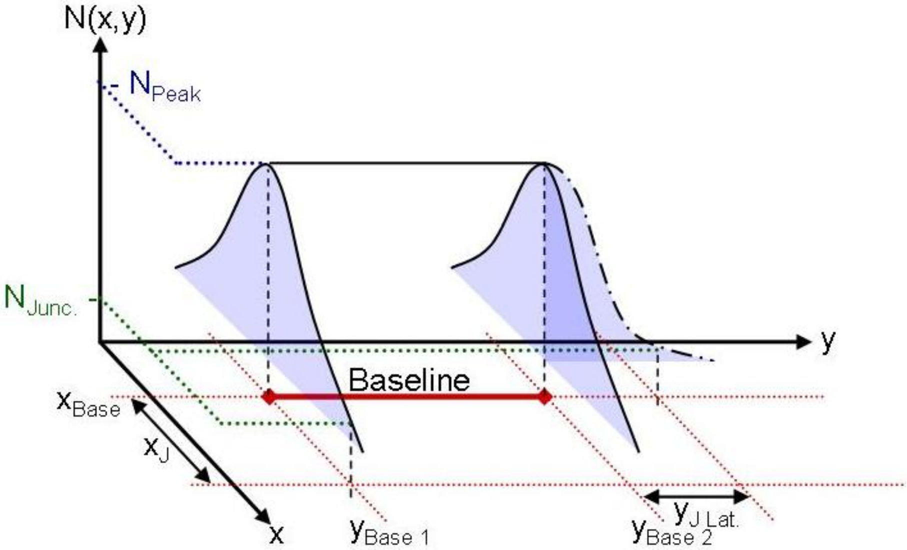

Figure 5-2: Illustration of the necessary definition parameters for a Gaussian Profile

All other profiles are implemented to represent doping profiles, based on high energetic ion implantation in combination with post annealing, being best represented by Gaussian distributions. These profiles are defined by multiple parameters, as illustrated in Figure 5-2. To locate such a profile, a baseline has to be defined by setting the two endpoints ($x_{Base}/y_{Base\ 1}$ and $x_{Base}/y_{Base\ 2}$). The peak value N_{Peak} defines the doping maxima, and is only present at the location of the baseline. In the direction perpendicular to the baseline, the doping follows a Gaussian profile (5-1).

$$N(x) = N_{Peak} \cdot \exp\left(-\frac{1}{2}\left(\frac{x - x_{Base}}{\sigma} \right)^2 \right) \tag{5-1}$$

With x_J being defined as the distance between the baseline and the location where the doping level dropped to N_{Junc} , the standard deviation σ is defined as follows:

$$\sigma = x_J \cdot \frac{1}{\sqrt{2 \cdot \ln\left(\frac{N_{Peak}}{N_{Junc.}} \right)}} \tag{5-2}$$

Exceeding the baseline in y- direction, the profile also shows a Gaussian dependence. The standard deviation in lateral direction is set by use of a factor $K_{Lat.}$, with $\sigma_{Lat.} = \sigma \cdot K_{Lat.}$, or equally delivering $y_{J\ Lat.} = y_J \cdot K_{Lat.}$. Especially for the Halo- and Ldd- implants, where the baselines end in close proximity to the channel region, $K_{Lat.}$ has a strong influence onto the resulting channel doping profile and will be one of the used fitting parameters for the model adjustment later on.

(4) Meshing Strategy

Since the mesh defines the local density of calculation points, it directly affects the quality and the process time of a simulation. Consequently, any mesh strategy is a compromise between process time and simulation accuracy. The used semiautomatic mesh engine defines the node density within user defined maximum and minimum spacing values, based on the local doping gradient. The enlarged view of an n-FET channel in the right part of Figure 5-2 shows the resulting mesh, whereas every intersection of the black lines represents one calculation point for the numerical solver. The minimum spacing is present at the surroundings of the gate oxide/channel interface with vertical distances below 0.1nm. But, since the mesh does not need to be varied throughout the later simulations, it will not be discussed in more detail.

5.1.2. Sentaurus Device

Based on the predefined device structure, the numerical simulation engine of Sentaurus Device can now solve the assigned set of equations to determine the charge distribution, terminal currents and voltages and also thermal effects. All DC simulations are done in a quasi static way by at first solving all assigned equations with all terminals set to 0 V to find an initial solution for a given structure. With the n-FET in Figure 5-2, the simulation would, for example, determine the expense of the source-, drain- and channel depletion regions, using the given boundary conditions until the solution exceeds a user defined convergence level. Based on the simulation task, the terminal voltages will then be subsequently ramped up to predefined values, whilst storing the desired information.

Regarding that the later simulations will only serve as additional support for the discussion of the experimental results, a full discussion of all included semiconductor equations (as can be found in the Sentaurus Device manual [24]) would by far exceed the scope of this work. Therefore, only the most important calculation models, which will in part be used for the model adjustment in the following subsection, will be briefly reviewed.

The key parameter for the correct simulation of the channel current is the local carrier concentration and mobility. The concentration is modeled with regards to quantum effects in the inversion channel, following a standard model. The carrier mobility is derived with much higher complexity. Following Mathiessen's rule, the resulting mobility can be calculated by:

$$\frac{1}{\mu}=\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}+\cdots+\frac{1}{\mu_n}\ , \tag{5-3}$$

where μ_1 to μ_n represent the different mobility contributions. With a MOSFET structure, there are in general four different systematic influences on the resulting mobility: the lattice temperature, the local doping- and carrier concentration, the electrical field normal to the SiO_2/Si interface and finally the high field saturation initiated by the parallel electric field. Three separate models are utilized to represent these physical effects and they will be discussed in the following:

(1) Lattice Temperature and Doping Concentration Dependence

The first two parameters are the lattice temperature and the doping- and free carrier concentration. The utilized model is the Philips unified mobility model, proposed by Klaassen [25]. In addition to describing the temperature dependence of the mobility, the model takes into account electron–hole scattering, screening of ionized impurities by charge carriers, and clustering of impurities. With the index i being either n for electrons or h for holes, the bulk mobility $\mu_{i,bulk}$ is calculated as follows:

$$\frac{1}{\mu_{i,bulk}}=\frac{1}{\mu_{i,L}}+\frac{1}{\mu_{i,DAeh}} \tag{5-4}$$

The lattice temperature dependence is given by:

$$\mu_{i,L}=\mu_{i,\max}\left(\frac{T}{300K}\right)^{-\Theta_i} \tag{5-5}$$

Hereby $\mu_{i,max}$ represents the maximum bulk mobility and Θ_i is a fitting parameter. The calculation of $\mu_{i,DAeh}$ is more complex, but since this model has not been modified at all (only used with default settings), it does not require further discussion.

(2) Normal Electric Field Dependence

The second model represents the normal electric field ($\xi_\perp$) dependent mobility degradation at an interface (gate oxide/channel). The enhanced Lombardi model in combination with coefficients derived by Lucent [26] was found to be most accurate. The resulting low field mobility $\mu_{i,low}$ is computed as follows:

$$\frac{1}{\mu_{i,low}}=\frac{1}{\mu_{i,bulk}}+\frac{D}{\mu_{ac}}+\frac{D}{\mu_{sr}} \tag{5-6}$$

Hereby D = exp(-x/l_{crit}) is a damping factor, supposed to switch off the inversion layer term far away from the interface. The model separates between the influences of acoustic phonon scattering (μ_{ac}) and surface roughness (μ_{sr}) with

$$\mu_{ac}=\frac{B}{\xi_\perp}+C\frac{\left(N_{tot}/N_0\right)^\lambda}{\xi_\perp^{1/3}\cdot\left(T/300K\right)^k} \tag{5-7}$$

$$\frac{1}{\mu_{sr}} = \left(\frac{\left(\xi_{\perp} \big/ \xi_{ref} \right)^{A^*}}{\delta} + \frac{\xi_{\perp}^{\ 3}}{\eta} \right)^{-1} \tag{5-8}$$

and

$$A^* = A + \frac{\alpha_{\perp}(n+p)N_{ref}}{\left(N_{tot} + N_1\right)^{v}} \ . \tag{5-9}$$

Besides the normal electrical field, other local quantities as the sum of all present doping atoms N_{tot} and the sum of all free carriers (n + p) are included for the calculation. All other default parameters are listed in Table 5-1, at the end of this subsection. The fitting factors λ and A will be used to adjust the simulation model to the measurement results of the utilized test structures.

(3) High Field Saturation (HFS)

The Carnali model [27] was assigned to cope for the velocity saturation in high field regions. In contrast to all previous mobility models, the high field saturation mobility is not computed by use of Mathiessen's rule, it rather calculates the final mobility by using all previous results as input. A basic version (5-10) shows the key parameters and allows for a better understanding of the underlying dependencies. With this model, the electric field as the driving force is derived by conventional transport equations (e.g. drift/diffusion model):

$$\mu\left(\xi_{hfs}\right) = \frac{\mu_{low}}{\left[1 + \left(\frac{\mu_{low} \cdot \xi_{hfs}}{v_{sat}} \right)^{\beta} \right]^{1/\beta}} \ , \tag{5-10}$$

with

$$\beta = \beta_0 \left(\frac{T}{300K} \right)^{\beta_{\exp}} . \tag{5-11}$$

But since the conventional transport equations tend to overestimate the resulting carrier velocity and impact ionization in high field regions of modern short channel FETs, a more accurate set of differential equations has been developed throughout the past two decades, referred to as the Hydrodynamic transport model. This model derives the energy of free carriers in form of a carrier temperature (T_n and T_h for electrons and holes respectively). Due to the calculation complexity, the Hydrodynamic model is only applied to the carrier type of the FET inversion charge (e.g. electrons for an n-FET). There have been numerous publications regarding model extensions and new parameter sets, all being listed in the Sentaurus Device manual for further reading [24].

For the aim of local mobility simulation, the carrier temperature has to be transformed into an appropriate driving force for the high field saturation model as follows

$$\xi_{hfs} = \sqrt{\frac{\max\left(3/2 \cdot k(T_n - T);0\right)}{\tau_e q \mu}} \quad , \tag{5-12}$$

with T being the local lattice temperature and τ_e the energy relaxation time. Substituting equation (5-12) into (5-10) yields the Hydrodynamic Canali model:

$$\mu\left(\xi_{hfs}\right) = \frac{\mu_{low}}{\left[\sqrt{1+\gamma^2 \max\left(3/2 \cdot k(T_n - T);0\right)^{\beta}} + \gamma \max\left(3/2 \cdot k(T_n - T);0\right)^{\beta/2}\right]^{2/\beta}} \tag{5-13}$$

where γ is given by:

$$\gamma = \frac{1}{2}\left(\frac{\mu_{low}}{q \cdot \tau_e \cdot v_{sat}^2}\right)^{\beta/2} . \tag{5-14}$$

The default parameters are listed in Table 5-1. According to this model, the fitting parameters β_0 and v_{sat} will later be used for adjustment.

Table 5-1: Default coefficients for the used E-normal and HFS models

Symbol	Electrons	Holes	Unit
B	$3.61 \cdot 10^{7}$	$1.51 \cdot 10^{7}$	cm / s
C	$1.7 \cdot 10^{2}$	$4.18 \cdot 10^{3}$	$cm^{5/3}V^{-2/3}s^{-1}$
N_0	1	1	cm^{-3}
λ	0.0233	0.0119	1
k	1.7	0.9	1
δ	$3.58 \cdot 10^{18}$	$4.1 \cdot 10^{15}$	cm^2/Vs
A	2.58	2.18	1
$\alpha_\perp$	$2 \cdot 10^{-20}$ *	$3 \cdot 10^{-20}$ *	cm^{-3}
N_{ref}	1	1	cm^{-3}
ξ_{ref}	1	1	V/cm
N_1	1	1	cm^{-3}
v	0.0767	0.123	1
η	$5.82 \cdot 10^{30}$ *	$2.0546 \cdot 10^{30}$ *	$V^2cm^{-1}s^{-1}$
l_{crit}	1	1	cm
β_0	1.109	1.213	1
β_{exp}	0.66	0.17	1
v_{sat}	$1.07 \cdot 10^{7}$	$8.37 \cdot 10^{6}$	cm/s

(* indicates differing parameters compared to the Lucent model, modified to correct for changes associated with the implemented quantum correction in the inversion channel, based on an example model of the Sentaurus Device library - CMOS Characterization [24])

5.2. DC Transistor Model

The device simulations will be used to support experimental results, all measured on 120 nm technology test structures, meaning that the nominal layout gate length (L_{layout}) is 120 nm. The first experimental part will focus on the FIB modification of single MOS-FET structures but prior to that, the necessary simulation models have to be derived. The layout of one n-FET is shown in Figure 5-3. All test transistor terminals (gate, source & drain) are accessible via separated terminals. The transistors have 10 µm width and the drain length can be extracted to $L_d = (L_{total} - L_{layout})/2 = 1.76$ µm.

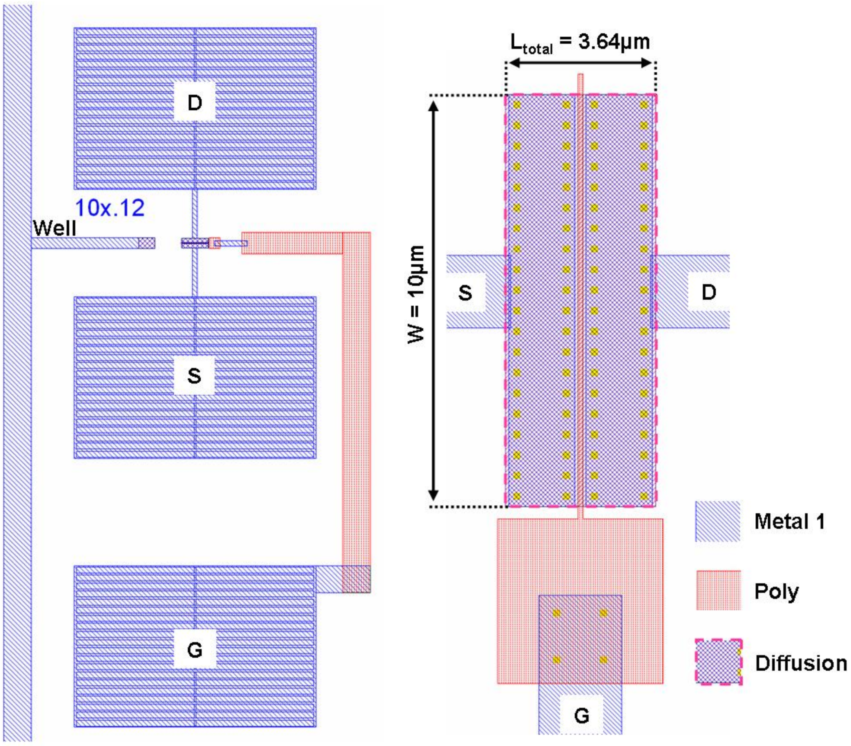

Figure 5-3: Layout of n-FET test structures: overview (left) & detailed image (right), rotated by 90°

The most important geometry parameter is the real physical gate length (L_{gate}). Figure 5-4 shows an SEM image of a cross section through an n-FET with $L_{layout} = 140$ nm. The measurement of the resulting physical gate length shows a shrink of ≈15 nm, similarly present at all transistors of the same technology node. Consequently, the physical gate length of the simulated transistors is set to $L_{gate} = L_{layout} – 15$ nm, for all simulations. Furthermore, the width of the drain/source area is only of minor importance for the DC simulations and can be rounded to $L_d = 1.8$ µm.

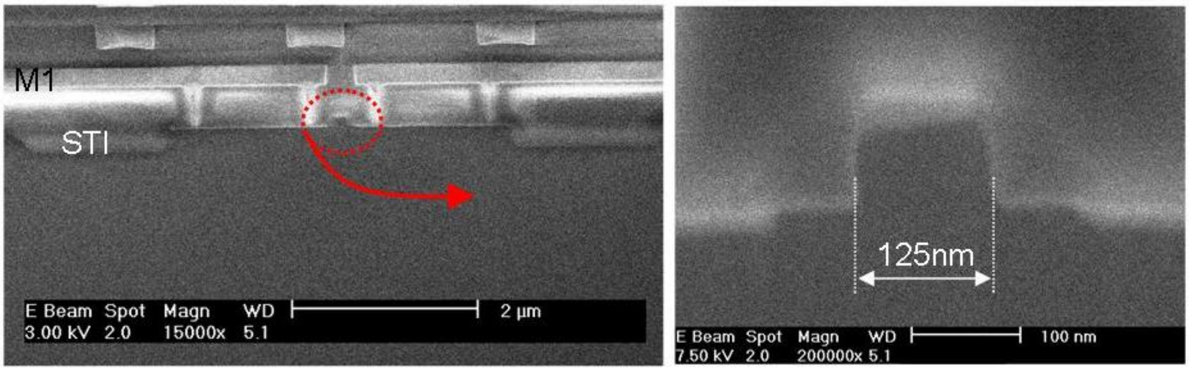

Figure 5-4: Cross section of an n-FET, fabricated in 120 nm technology, with L_{layout} = 140 nm

The implementation of the necessary doping profiles was done in accordance to secondary ion mass spectroscopy (SIMS) measurement results of the same technology, and will be shown with the single models for the n- and p-FET later on. But due to the nature of SIMS, meaningful doping profiles can only be measured on horizontally constant implants on wide areas. Referring to the definition of a doping profile within the simulation environment (as illustrated in Figure 5-2), SIMS results do not allow to determine the lateral expanse of a profile, exceeding its baseline. Hence, the parameter $K_{Lat.}$ remains undefined. As illustrated in Figure 5-5, this undefined lateral factor is of major importance for the source/drain-, the halo- and the Ldd implants, defining the resulting ΔS/D, ΔLdd and Δhalo.

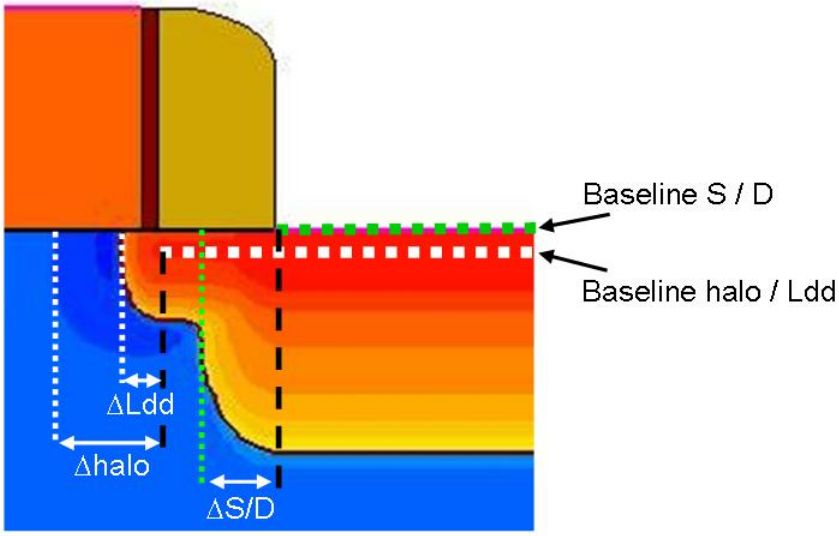

Figure 5-5: Illustration of the lateral expanse of the S/D, halo and Ldd doping implants

The source/drain profile is modeled in a way that the implant ends clearly before the gate region, which was not varied throughout the model adjustment. The junction between the Ldd and halo implant defines the metallurgical gate length (L_{met}). Consequently, the quality of the later simulation model is greatly dependent on the right balance between Ldd and halo, to set the right threshold voltage and to control short channel effects like DIBL.

The well-contact resistance is one key parameter for the simulation, whereby the implementation into the simulation model is not trivial. With the real test structures (see layout in Figure 5-3), there is only one well-contact in approximately 5 µm distance to the end of the gate. Since all simulations were done in 2D mode, a simplified geometry had to be generated, delivering equivalent electrical behavior with reduced dimensions, hence an acceptable number of mesh nodes and resulting simulation time. Figure 5-6 shows the final structure. Two well-contacts were established on both sides in 1.5µm lateral distance to the end of source and drain. The STI shapes were rounded to suppress edge effects and the meshing distance was reduced at the Si/STI interface and at the well-contacts to gain more accurate results and a higher numerical stability. The well-contact resistance was then scaled with t_{Si} in accordance to the extracted R_{well} values of the real FETs. Since two parallel well-contacts are declared in the simulation (well-contact right and left), the individual value always had to be twice the measured resistance. In the following, the reference devices are all simulated with $R_{well\ left} = R_{well\ right} = 20\ \Omega$ in parallel, hence $R_{well} = 10\ \Omega$.

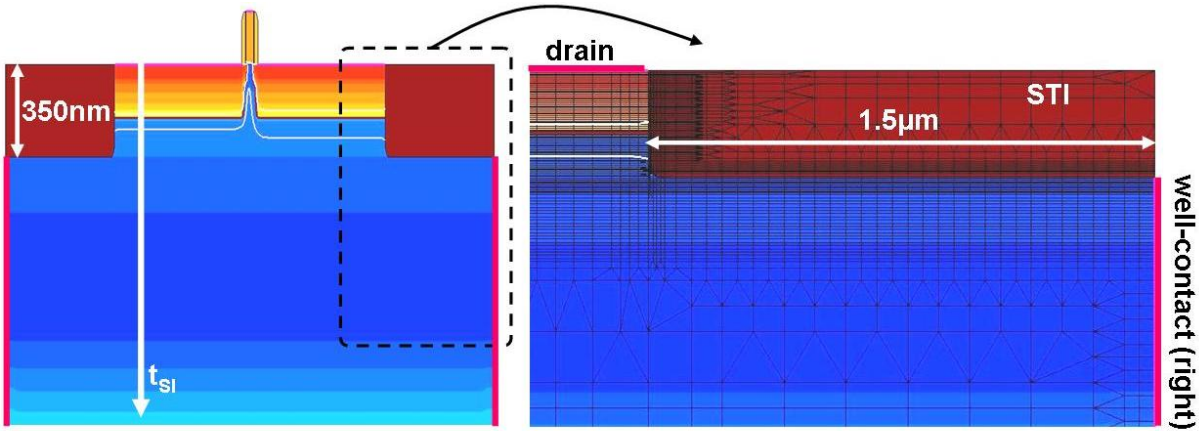

Figure 5-6: Structure of FET simulation model with enlarged view of the right STI and well-contact including the utilized mesh

5.2.1. n-FET DC Reference Model

The DC model adjustment was carried out in two main steps: firstly the implementation and alignment of the doping profiles and finally the adjustment of the utilized calculation model parameters.

As discussed earlier, the vertical doping profiles could be implemented with regards to SIMS measurements (Figure 5-7). The upper plot shows the cross-sectional view of the resulting drain/source doping, whereas the lower reflects the channel and well-doping.

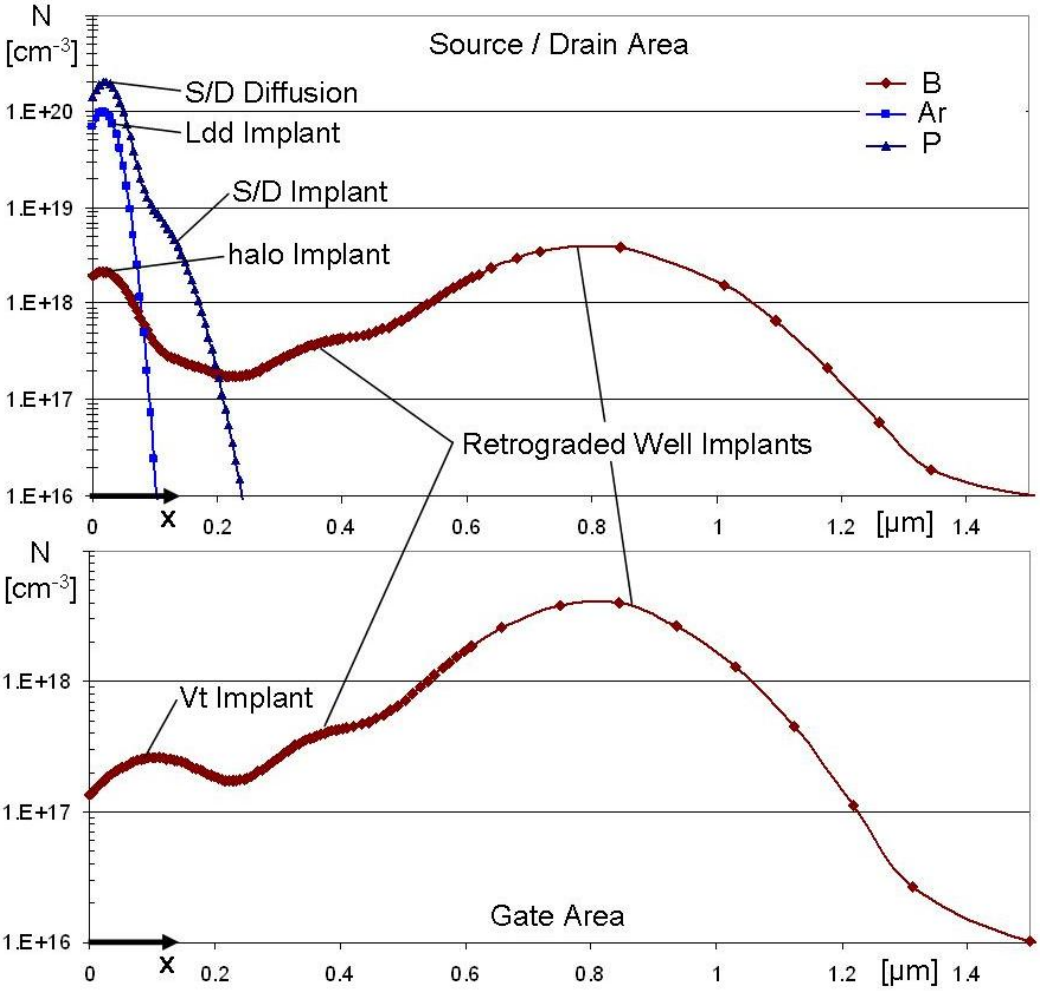

Figure 5-7: Doping profiles of the simulated n-FET, shown as a cross-sectional measurement through the drain (upper) and the gate (lower)

The progress of the model adjustment is illustrated in Figure 5-8. Based on the vertical doping profiles, similar characteristics to (A) were simulated. In the next step, the lateral (or horizontal) expanse of the halo and Ldd implants were adjusted to match the linear V_t (measured with $V_{DS} = 0.05$ V) as well as V_t measured with $V_{DS} = 1.2$ V, regarding the real transistor data (B). In this way, short channel effects as DIBL are represented correctly in the simulation model.

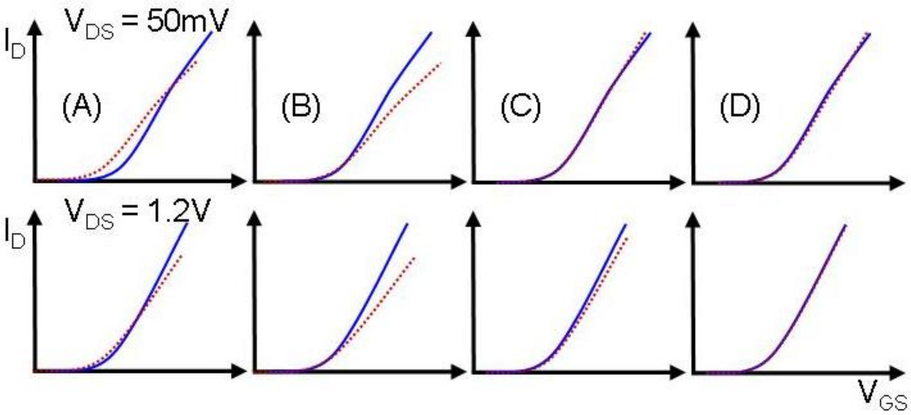

Figure 5-8: Illustration of transfer characteristics with V_{DS} = 50 mV (upper row) and 1.2 V (lower row), showing the changes associated with the subsequent model alignment steps (blue = reference 120 nm FET, red = simulation results)

From now on, all doping profiles are kept the same and the FET characteristics are aligned by tuning the fitting parameters, identified in 5.1.2. First, the mobility dependence to the normal electric field is modified by use of the parameters λ and A (see formula 5-7). With V_{DS} being set to 50 mV, the drain current progression is not affected by high field saturation effects, allowing finding the right λ and A without other disturbing effects (C). After that, the maximum drain current was adjusted by modifying the high field saturation parameter β_0 (see formula 5-11) as illustrated in Figure 5-8 (D).

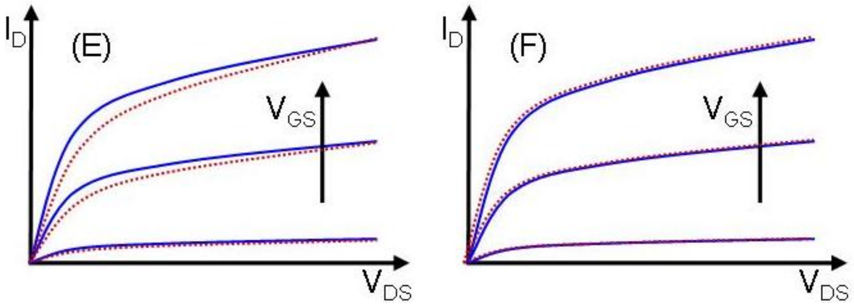

Figure 5-9: Final model adjustment, aligning the slope in saturation of the output characteristic

Finally, the slope in the saturation region of the output-characteristic (known as output conductance g_D) had to be adjusted by reducing v_{sat} (see formula 5-14), with the influence being illustrated in Figure 5-9.

All DC simulations also include self-heating effects. The resulting local temperature is based on the balance between local power absorption and thermal flux, and is greatly influenced by the thermal boundary conditions. In most standard simulation approaches for bulk MOS FETs only the bulk is considered as heat sink. But due to the later thinning of the investigated bulk FETs, gradually turning them into SOI like structures, the thermal boundary conditions needed to be implemented with more care, which are discussed in combination with the experimental results in chapter 6.3.

With regards to the later use of the adjusted FET models, the characteristics were adjusted with priority on the saturation transfer characteristic (V_{DS} = 1.2 V) and the output characteristic, since they dominate the dynamic performance of a CMOS inverter. Especially these characteristics are in excellent agreement to the real transistor data as shown in Figure

5-10. Also the linear transfer characteristic matches according to $V_{t\ lin}$ and I_D for $V_{GS} < 1$ V, whereby the maximum mismatch remains below 8% with $V_{GS} = 1.2$ V.

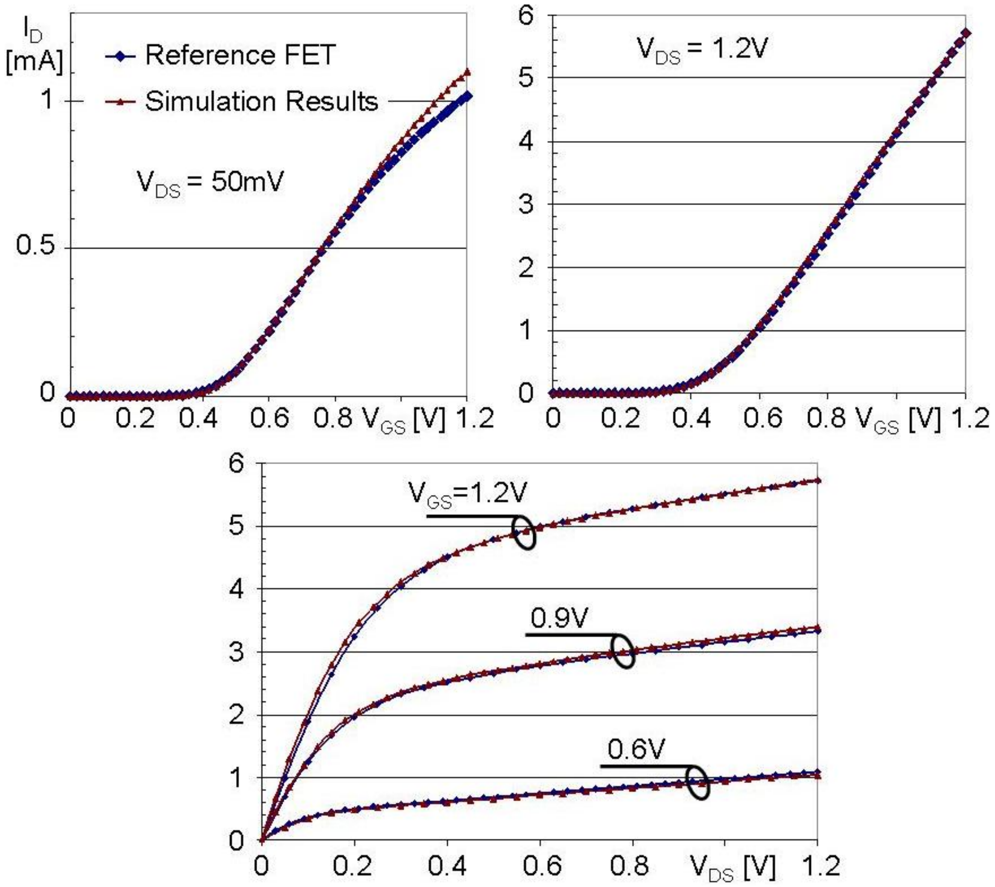

Figure 5-10: Reference device characteristics compared to simulation results for the 120 nm n-FET

5.2.2. p-FET DC Reference Model

The p-FET model was derived in the same way as described for the n-FET. Figure 5-11 shows the implemented doping implants. In contrary to the n-FET, the p-FET is fabricated with two channel (or V_t) implants, once using Arsenic, once Phosphorus as dopant. Since the Ldd- and S/D implant are both Boron doped, the Ldd profile can not be separated in a SIMS measurement and the implant properties were estimated to have the same peak value and distribution as the n-FET Ldd.

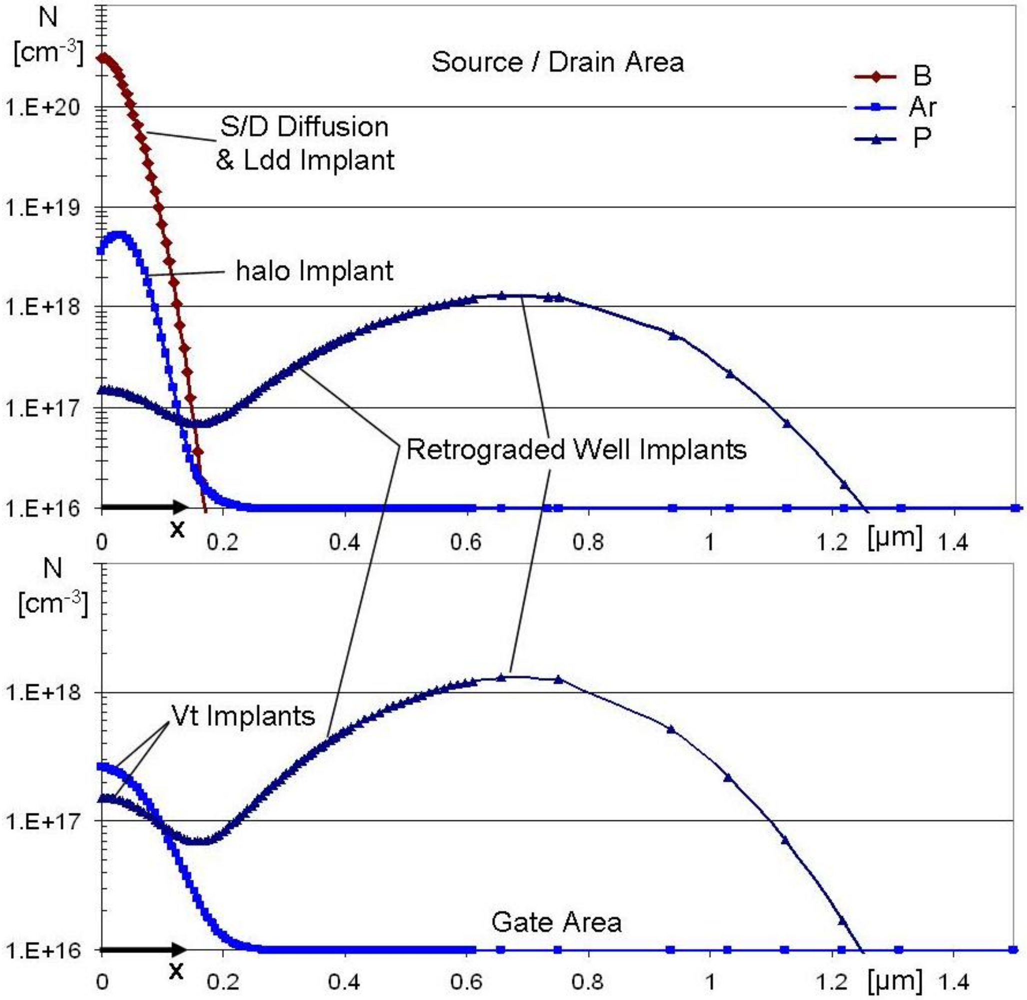

Figure 5-11: Doping profiles of the simulated p-FET, shown as a cross-sectional measurement through the drain (upper) and the gate (lower)

The resulting DC characteristics of the p-FET are plotted in Figure 5-12, and are compared to real device data. All thee characteristics are in excellent agreement with the reference device data. The necessary changes in the earlier described fitting parameters are listed in Table 5-2.

An exemplary n-FET structure definition file, the execution command file for a dynamic simulation and the finally used Silicon parameter file can be found in appendix B.

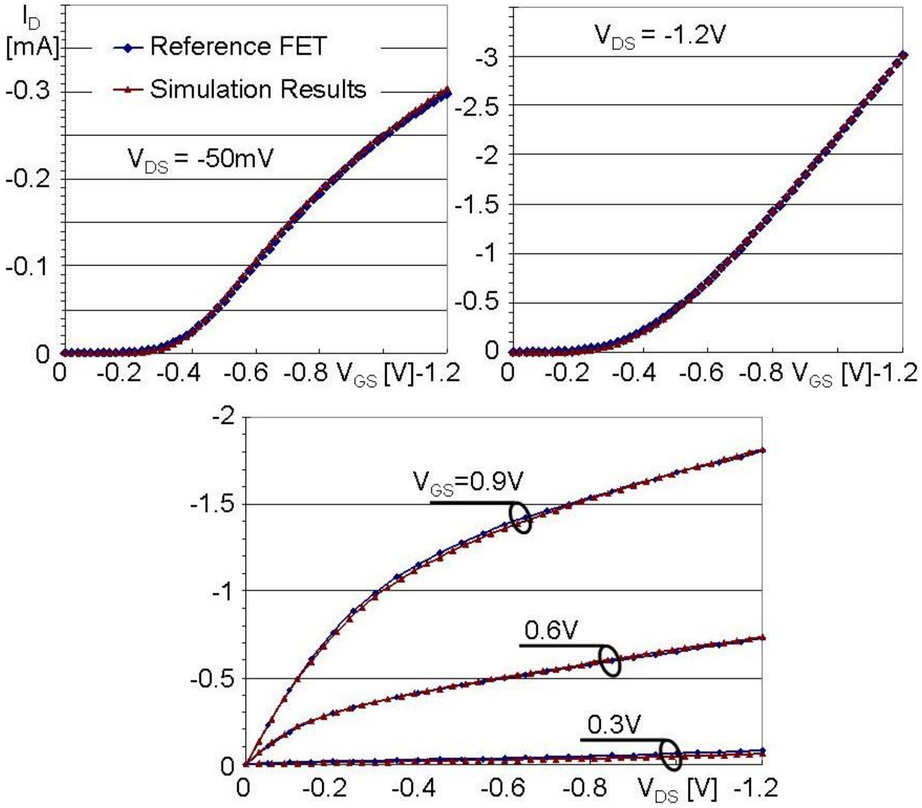

Figure 5-12: Reference device characteristics compared to simulation results for the 120 nm p-FET

Table 5-2: Summarized fitting parameter values for the adjustment of the n- and p-FET

Symbol	Electrons / n-FET default => new	Holes / p-FET default => new	Unit
λ	0.0233 => 0.077	0.0119 => 0.0158	1
A	2.58 => 2.582	2.18 => 2.118	1
β_0	1.109 => 2.02	1.213 => 2.25	1
v_{sat}	10.7 => 9.1	8.37 => 8.13	10^6·cm/s

5.3. Dynamic Device Simulations

The dynamic simulations are used to evaluate dynamic device performance, compared to measurement results on different ring oscillators undergoing the FIB treatment. Due to simulation complexity, it is impossible to simulate a full oscillator comprised of 65 single inverters at the same time. But the simulation environment allows combining the above derived single FET models to one CMOS inverter, where the n- and p-FET can be simulated simultaneously. Therefore, to drive the inverter similar to one of the RO gates, the oscillating signal of the real oscillator ring has to be reproduced and applied as an input signal. But before adjusting the dynamic simulation model, the general organization and working method of the dynamic simulation environment will be explained.

At first, the single FETs are generated separately by the Sentaurus structure editor, based on the same definition file as used for the single FET DC simulations. In the next step, the single devices are combined by use of a surrounding SPICE environment, connecting the relevant terminals of the transistors and the additional capacitive load ($C_{L\,sim}$) by assigning them to certain nodes, as illustrated in the compact model depicted in Figure 5-13. Parameters like transistor width, terminal resistances and also the set of relevant equations can be set for the n- and p-FET separately. Before the dynamic simulation can start, the static outer voltages have to be applied to the circuitry. For the depicted inverter, node 2 (n2) is constantly connected to V_{DD} and the input is set to 0. Consequently, the p-FET is on, passing V_{DD} to the common drain node (n3), the n-FET is off and $C_{L\,sim}$ is loaded to V_{DD}. Once the whole setup reaches a static equilibrium, a dynamic signal can be applied to the inverter input (n1), triggering the switching process.

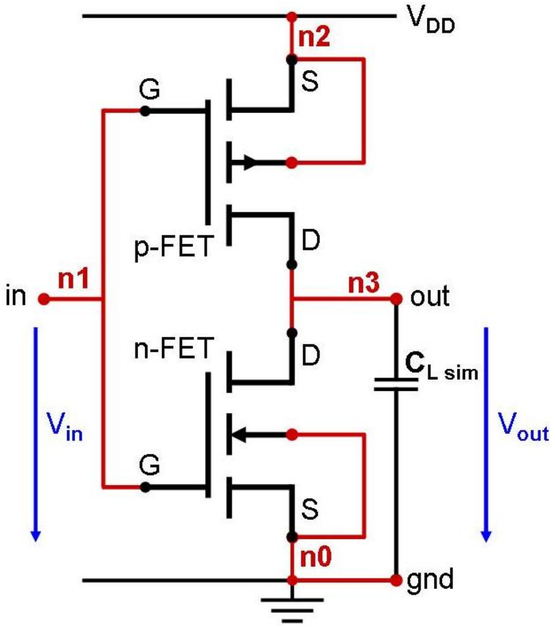

Figure 5-13: Compact model of simulated CMOS inverter, based on the n- and p-FET model plus a load capacitor (C_L), combined by four circuit nodes (n0 to n3)

6. Influence of Backside FIB Thinning on Static FET Performance

The proposed FIB thinning procedure resulting in ultra thin silicon is not only the basis for backside CE innovations but also allows for novel application methodologies in the field of optical- or particle-beam related circuit analysis and stimulation tools, which will be discussed in the chapters 9 and 10. Consequently, it is fundamentally important to know about the invasiveness of the initial thinning. This chapter will discuss the influences of the proposed FIB procedure on FET DC performance, starting with a general discussion of the expected alterations, followed by experimental results combined with simulations.

6.1. Expected Influence on Static Device Performance

The UtS formation process influences several device parameters. Figure 6-1 illustrates three different stages of the FIB procedure, all attributed with having different impact onto the modified device.

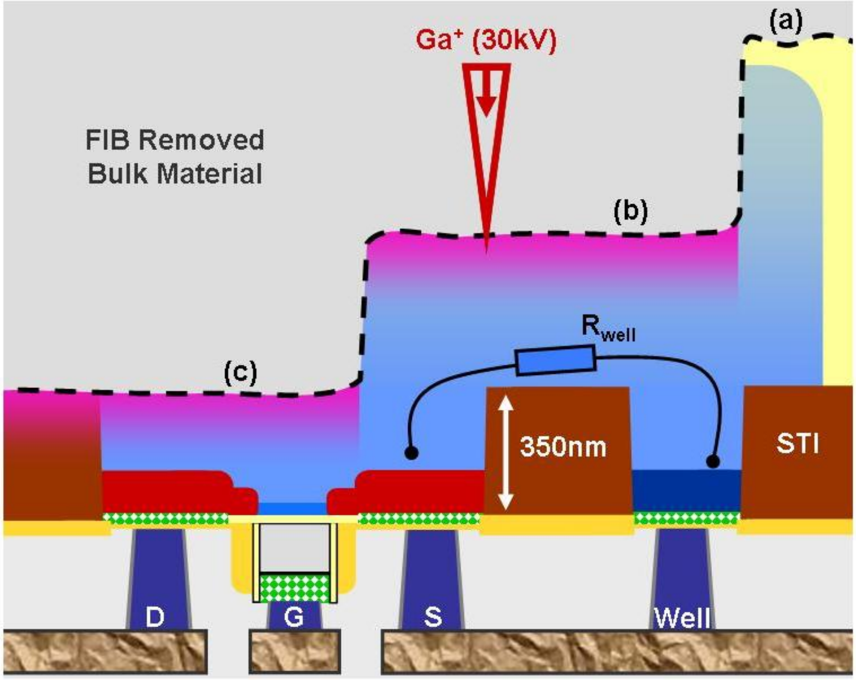

Figure 6-1: Illustration of a FIB-thinned p-FET

Any backside CE process begins with mechanical thinning, followed by coarse FIB milling, locally removing the bulk in the area of interest. This process is stopped on n-well level (a) with a remaining Si thickness of 1.5-3 µm. Assuming an initial wafer thickness of a few hundred micrometers, these preparation steps already removed almost the bulk material. But since the intrinsic device, the channel, the drain/source areas and the majority of the well- or body material are not affected yet, nothing but the thermal management / self-heating of the circuitry is expected to be altered.

For the continuation of the process, the local damage related to the 30 keV Ga^+ ion bombardment of the rear surface has to be taken into account. Various simulations and experiments (e.g. [14] & [15]) have shown that a 30 keV Ga^+ ion beam creates an amorphous top layer of ≈ 30 nm thickness. But also the material underneath this amorphous top layer will suffer from degradation, having an increased number of crystalline defects combined with Gallium contamination. The affected area is marked by the pink gradient in Figure 6-1 and will show a strongly decreased carrier mobility (or conductivity) and minority carrier lifetime.

In the next step, the FIB thinning reduces the remaining n-well thickness, gradually increasing the lateral well-resistance R_{well}. As a result, the remaining FET body will become more and more isolated, finally being almost electrically floating. At an intermediate thickness (b), the increased well-resistance will initiate SOI like effects, observable as a kink in the output characteristic. The progression of the well-resistance can be monitored by measuring the forward biased drain to well diode throughout the thinning process.

Below a certain thickness, the ion beam related damage is expected to start degrading the lateral source/drain to well junctions. The additional crystalline defects and Ga contaminants will start interfusing the depletion layer. These additional imperfections act as additional generation centers, increasing the diode reverse current.

Finally, also the channel will be affected, resulting in decreased channel mobility and an overall decrease of the transistor performance.

6.2. Measurements on Single 120 nm Transistors

The first set of experiments was carried out using single MOS-FET structures, all fabricated in 120 nm technology, meaning that L_{layout} = 120 nm. The layout of one n-FET is shown in Figure 5-3. All test transistor terminals (gate, source & drain) are accessible via separated bond pads, except for the common well-contact, which is shared by ten transistors.

To allow for the desired measurement accuracy, all later data were measured in-situ, directly in the FIB vacuum chamber, following to each step of the FIB thinning procedure. Consequently, the measurement environment could remain unchanged throughout the full experiment, eliminating the negative influence of setup variations (mainly package to burn-in-socket resistance and surface reactions in atmospheric environment). The lower current limit of the utilized vacuum setup A (appendix A) was found to ≈ $1 \cdot 10^{-10}$ A. The largest, 10µm wide transistors were chosen for this experiment to minimize the measurement error induced by this setup limit.

To allow a more intuitive understanding, all measured data will be presented as a function of the remaining silicon (or device) thickness (t_{Si}). But regarding the flow of the FIB thinning procedure, it is impossible to accurately measure this thickness directly during the process since a precise measurement would require the fabrication of a cross-section combined with tilting of the device, being destructive and also incompatible with the used vacuum setup. To overcome this systematic problem, only three direct thickness values have been extracted by use of cross-sections after the preparation was ended and all intermediate steps have been calculated by means of linear interpolation over the etch-time.

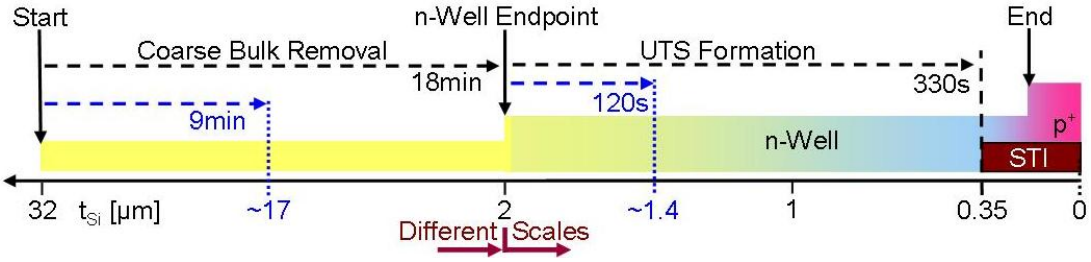

Figure 6-2: Illustration of the used Si thickness calculation strategy.

The underlying interpolation points are the start thickness, the thickness on n-well endpoint level and the final device thickness after finishing the experiment (all determined by individual cross-section measurements of each sample/test-structure). Additionally, the technology based STI height is always 350 nm which could be used as an additional interpolation point (number four). Assuming a linear etch-rate, the time dependent thickness can be calculated as illustrated in Figure 6-2.

In the thickness range from start to n-well level, the FIB milling underlies severe process variations and the linear approximation can only give a rough estimate and may be off by several micro meters. But as discussed later, the influence of the FIB procedure on device performance is negligible in this thickness range. Therefore, these values will not be used during the later analysis and the level of accuracy can be considered uncritical.

Furthermore, all thickness values have to be understood as being the average thickness in the area of interest – here, the active transistor area, approximately 10x4 μm^2 – whereby the thickness variation across this area drastically depends on the quality of the process. For all later results on the single FETs, the final thickness variations were calculated to be below 30 nm on STI level, based on the time difference between first and full STI exposure in combination with the etch rate.

During the final thinning to STI and beyond, all process parameters were strictly kept the same to allow for highest accuracy of the linear approximation method. A basic characterization of the utilized FIB gas operation, as applied during the final phase of the experiments ($t_{SI} > 2$ µm, below n-well level), did show minor nonlinearities when applied for different periods of time (influence of setup time and Iodine gas flux). Consequently, the accuracy of the determined average device thickness is limited to approximate ±15 nm.

6.2.1. The p-MOS FET DC Performance

Since the p-FET is less susceptible to impact ionization, it shows less SOI specific effects and allows for an easier evaluation of thermal effects and junction- and channel-degradation, compared to the n-FET. The following plots will show DC characteristics, all measured at the same p-channel transistor (the layout is equivalent to Figure 5-3) undergoing the FIB thinning procedure. Note that the proposed FIB procedure ends with revealing the STI shapes at a thickness of 350 nm, whereas the shown results include measurements for even thinner structures. These additional data points have been collected to investigate the process margin for the UtS fabrication.

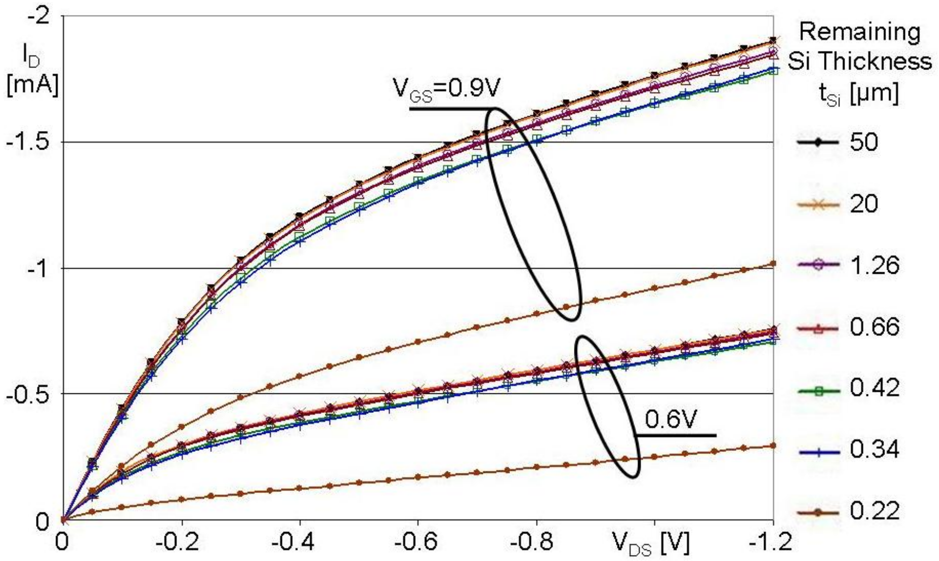

Figure 6-3: Output characteristics of a 10µm x 120nm p-FET as a function of silicon thickness

For all thicknesses down to 0.34 µm, the output characteristics in Figure 6-3 show only a small maximum current reduction (≈ 6.5%) with the subsequent bulk removal. Besides that, no SOI like effects (kink) are visible. Only a further material reduction clearly below the STI thickness strongly degrades the transistor performance, decreasing I_{Dmax} by almost 50% with t_{Si} = 220 nm.

Regarding a classical backside circuit edit, where only small parts (< 50%) of some single FETs are thinned to the STI to benefit from the STI alignment accuracy, these results already confirm the full functionality of the modified devices with uncritical performance degradation for most CE applications.

But for the use of backside analysis tools on the UtS platform, bigger areas covering multiple full transistors may be thinned to STI level. As a result, also minor changes at single devices can add up and may critically influence the DuT performance, cause fails or lead to adulterated timing measurements. To cope with this and also to gain a deeper understanding of the degrading effects, all changes in the various DC characteristics will be analyzed in the following, starting with the evaluation of I_{Dmax} and $V_{t\ lin}$.

Figure 6-4 shows several linear transfer characteristics, measured at different Si thicknesses. Based on this data, the linearly extrapolated threshold voltage ($V_{t\ lin}$, see chapter 4.1.1) has been extracted and is shown in the upper area of Figure 6-5, reading out to the left and drawn as a function of t_{Si}.

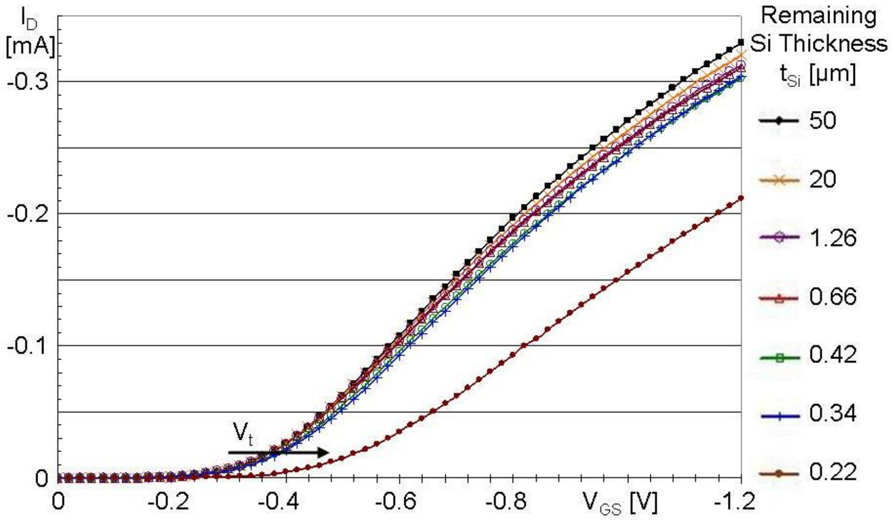

Figure 6-4: Linear transfer characteristics (V_{DS} =50mV) as a function of silicon thickness

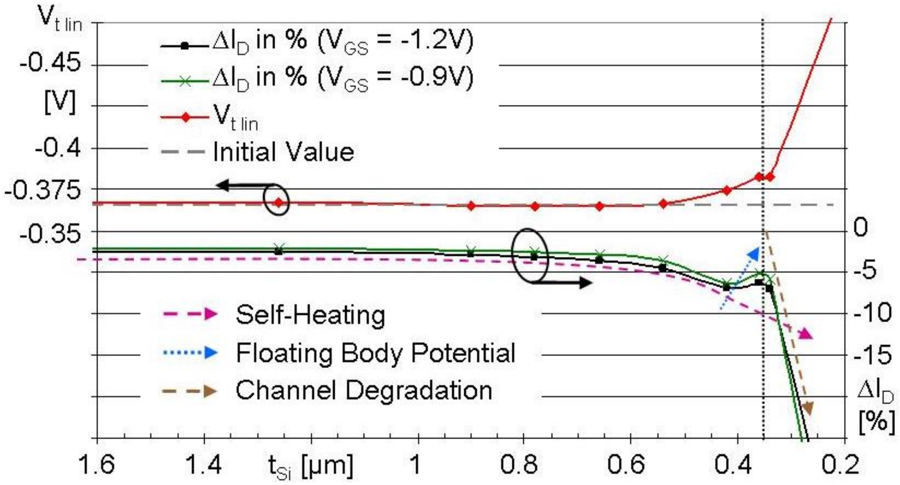

Figure 6-5: I_{Dmax} and $V_{t\,lin}$ as a function of silicon thickness, compared to initial values at t_{Si} = 50 µm

Over the full range of the proposed FIB thinning procedure (down to STI, with $t_{Si} \approx 350$ nm) $V_{t\,lin}$ changes by less than 20 mV. $V_{t\,lin}$ is extracted from measurements with V_{DS} = 50 mV and is consequently not affected by either impact ionization (ii) or temperature effects. $V_{t\,lin}$ rather reflects the intrinsic device parameters like channel-doping and is also very susceptible to body potential modulations/body bias. But the suspected reason for the $V_{t\,lin}$ increase will be discussed a little later in this chapter. Figure 6-5 also shows the degrading of I_D, plotted in percent regarding the initial value at t_{Si} = 50 µm. Both curves (V_{GS} = -0.9 & -1.2 V) were extracted with V_{DS} = -1.2 V and show a slow and steady current decrease (pink dashed line) which can be attributed to increased self-heating resulting from the bulk material removal. This is supported by the fact that the on-current decrease at V_{GS} = 0.9 V is smaller compared to V_{GS} = 1.2 V, driving only about 60% of the maximum current, hence facing less self-heating. The later simulations will allow a more quantitative discussion of the self-heating. Only as t_{Si} approaches STI height, this trend is shortly abandoned and the curves show a small local increase. At this point, the only possible reason is an increased body potential, reducing V_t in saturation (as can be seen from $V_{t\,lin}$, this does not happen in the linear region with V_{DS} = 50 mV). But since ii is unlikely to happen in a p-FET channel and also the output characteristics do not show a kink, the increased body potential will have to be explained after the discussion of the following drain to well diode characteristics. Continuing the FIB thinning below STI level drastically increases $V_{t\ lin}$ and is expected to degrade the channel mobility, altogether decreasing I_{Dmax} by more than 40% at t_{Si} = 220 nm.

To monitor the connectivity of the FET body to the outer well contact, drain to well diode characteristics have been measured subsequent to all FIB process steps (Figure 6-6). As discussed at the beginning of this chapter, the in-situ measurement equipment does not allow resolving currents below 100 pA, prohibiting correct in-situ measurements of the initial diode reverse current. A reference diode curve of a FET with exact same layout, from the same lot, has been measured on a bare die by use of a probe station in combination with a dedicated low current measurement setup (lower limit ≈ 1 fA, with the combination of Keithley 4200-SCS [28] and Karl-Süss PM8).

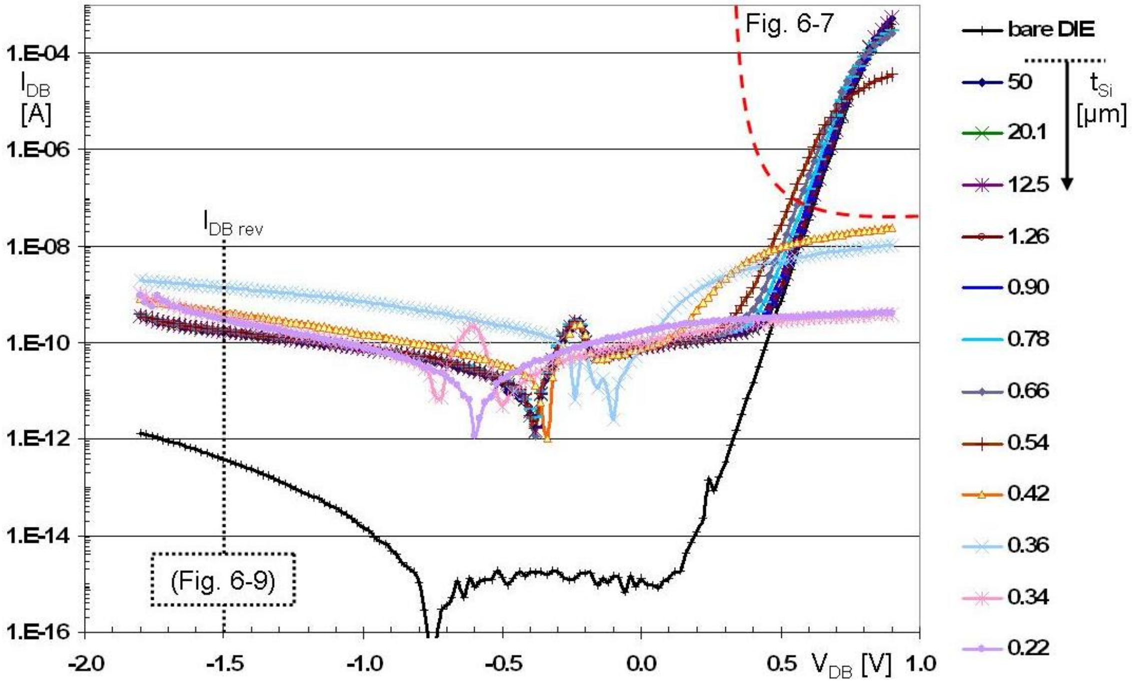

Figure 6-6: Drain to well diode characteristics as a function of silicon thickness, including a reference curve measured on a bare DIE

Especially the diode current in forward direction shows a strong alteration with the FIB thinning. As expected, the series resistance of the well-contact increases the thinner the remaining well-material becomes. Figure 6-7 shows an enlarged view of the forward direction in semi logarithmic scale, including pure logarithmic approximations of the diode currents for t_{Si} = 50 & 0.66 µm. Measuring the voltage difference between the approximation (straight lines) and the real diode current (ΔV_{DB}) allows extracting the well-resistance R_{well} to:

$$R_{well} \approx \frac{\Delta V_{DB}(t_{Si})}{I_{DB\_\max}(t_{Si})} \qquad (6\text{-}1)$$

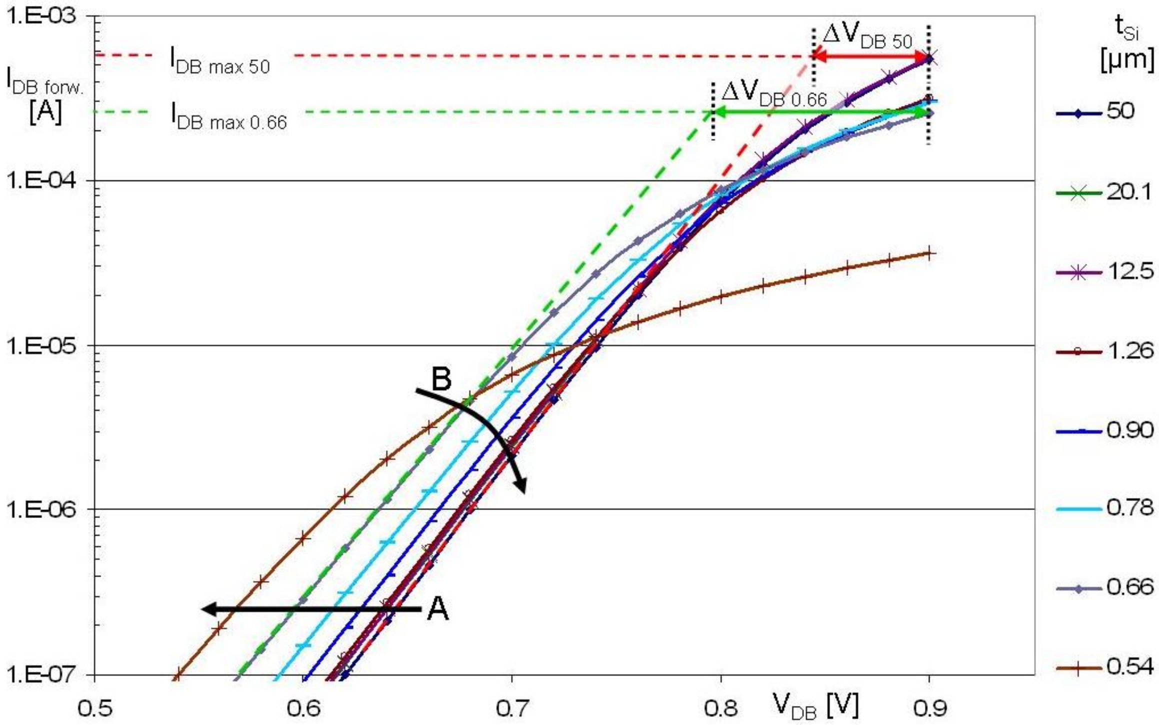

Figure 6-7: Enlarged view of the diode forward current as a function of silicon thickness

For all curves shown in Figure 6-7, a tangent can be applied in the voltage range where the curves show the ideal diode characteristic and appear linear in log scale. Only for the diodes measured with thicknesses below 0.54 µm, the curves hardly show any pure exponential region anymore, which does not allow extracting ΔV in the above described way. For a rough estimate, the resistance was calculated by constantly setting ΔV = 0.5 V, potentially underestimating the resistance values with $t_{Si} < 0.54$ µm.

The origin of the subsequent shift to lower voltages (A) and the additional slope reduction of the linear approximations (B) as marked in Figure 6-7 will be discussed later on.

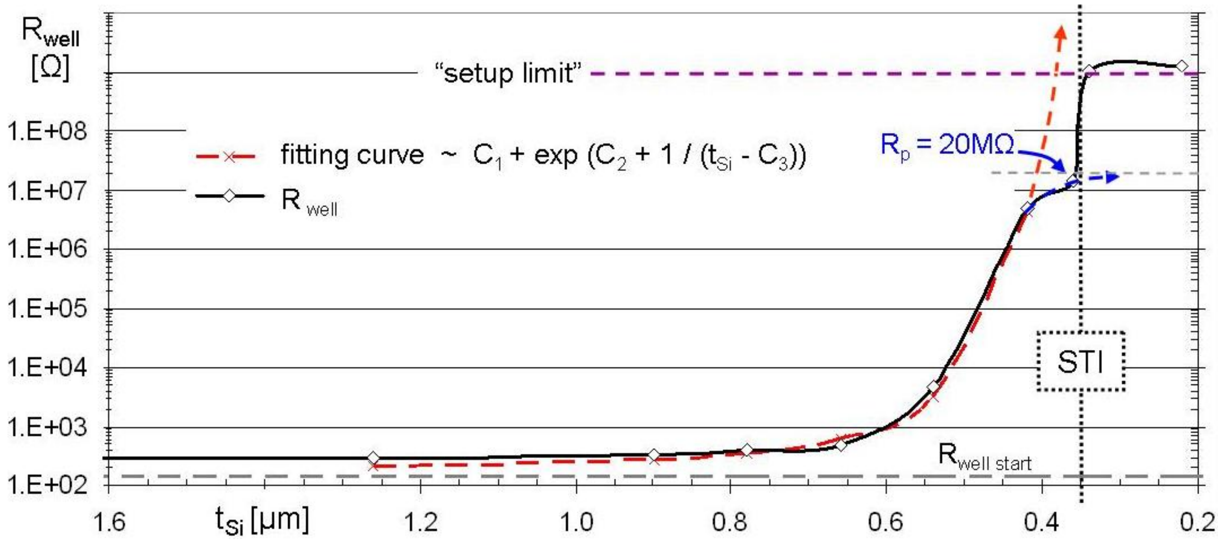

Figure 6-8: Calculated well-resistance and a fitting curve ($C_1 = 150$, $C_2 = 3$, $C_3 = 0.34$ µm) as a function of Silicon thickness

The progression of R_{well} is plotted in Figure 6-8. The initial value at t_{Si} = 50 µm was determined to be ≈ 150 Ω, shown as the dashed grey line. With the p-FET, only the n-doped well-material and not the rest of the p-substrate contributes to the ohmic well-connection. Consequently, R_{well} remains almost constant until t_{Si} approaches the well depth. The calculated values increase slowly from 156 Ω at t_{Si} = 12.5 µm to ≈ 470 Ω at 660 nm. Beyond this thickness, R_{well} increases dramatically and exceeds 1 GΩ with t_{Si} < 350 nm.

For thicknesses above 400 nm, the extracted R_{well} values could be fitted using an exponential formula, with $1/t_{Si}$ as argument, drawn as the orange dashed line. The exponential dependency can be explained by the retrograded well-doping profile (see Figure 5-11), which decays exponentially in the thickness range 0.7 µm < t_{Si} < 0.35 µm. Consequently, the sheet resistance of the well-material increases exponentially. Combined with the known 1/x dependence of R, with x being the height of any homogeneous conductor material (in this case t_{Si} minus the height of the STI shapes), this yields the chosen fitting formula. The renunciation from the exp(1/x) trend for t_{Si} < 400 nm is believed to be due to a higher remaining conductivity, associated with a remaining amorphous and highly Ga contaminated Si on top of the STI. A resistance of R_p = 20 MΩ in parallel to the fitting curve reproduced the R_{well} value for this measurement point (dashed blue line). Thinning below STI height, the underlying diode current for the R_{well} extraction fell below the setup limit and the extracted R_{well} saturated at approximately 1 GΩ, whereby the real resistance may increase further.

The diode reverse current ($I_{DB\ rev.}$ at V_{DB} = -1.5 V) is plotted in Figure 6-9. Due to the dominant setup leakage current, only the data points with t_{SI} < 500 nm show a clear response to the FIB thinning. Assuming a stable influence of the setup, it would be possible to remove it by subtraction of the average leakage current (determined to be ≈ 170 pA for the thickness range of 50 µm > t_{SI} > 1 µm). The resulting curve ($I_{DB\ rev.}$ - avg. leakage) is believed to reflect the real reverse current increase for t_{SI} < 600 nm, but for greater thicknesses it is too noisy to allow for a real quantitative evaluation of the FIB induced diode degradation.

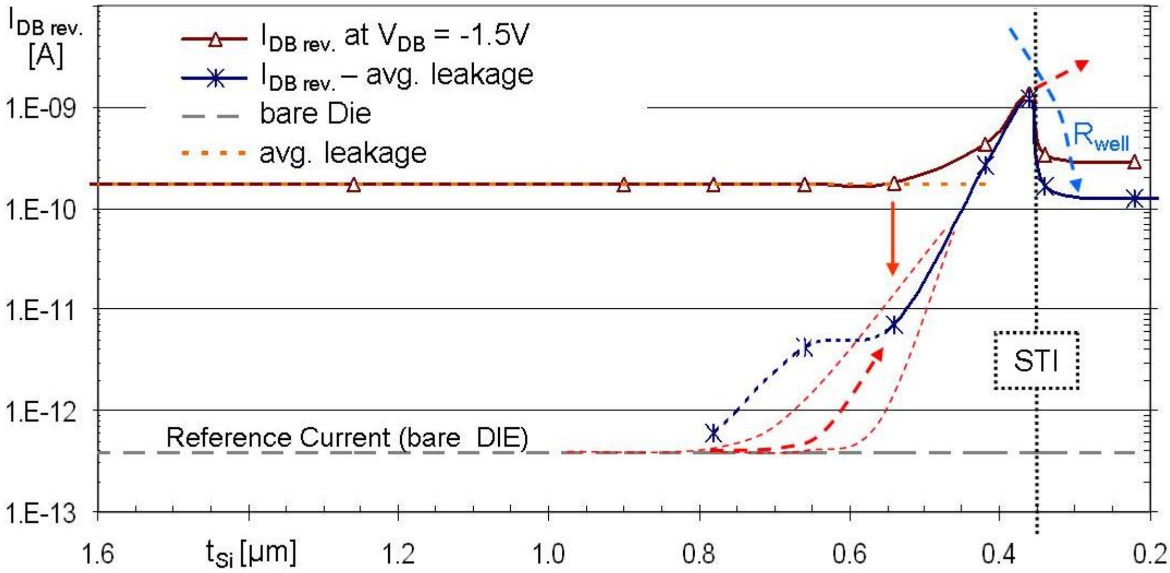

Figure 6-9 Diode reverse current at V_{DB} = -1.5 V as a function of Silicon thickness

The first clear increase at t_{Si} = 550 nm, shows already ≈ 20x reverse current increase. Based on the implemented doping profiles, the simulations show that the drain depletion region ends at t_{SI} ≈ 280 nm (below the gate-oxide level). Consequently, the ion beam must start degrading the reverse biased diode performance with a remaining distance of more than 270 nm between the surface and the end of the depletion region.

The FIB induced crystalline imperfections act as generation-, or recombination-centers, drastically reducing the average minority carrier lifetime in the affected volume. The additionally generated electron hole pairs can increase the diode saturation current as long as they are generated with less than a diffusion-length (L_{diff}) distance to the field (depletion-region) of the pn-junction. For the initial FET, without FIB induced degradation, L_{diff} would be ≥ 10 µm with the present well doping levels. But, L_{diff} also decreases similar to the carrier lifetime in the surface-near material of the FIB thinned DuT, explaining the measured reverse current increase.

The same effect applies for the increased forward current values and the reduced slope in forward direction of the diodes (marked as A and B in Figure 6-6 respectively). The well-contact is fabricated on device level (upper Si surface, due to the planar process technology) and the diode current (and associated recombination) is almost completely organized in the uppermost 1.5 µm of the transistor/well material. Consequently, the effect becomes visible at $t_{SI} \approx 1.2$ µm, long before the reverse current shows response to the FIB damage.

The precise investigation of the FIB-induced crystallographic defects, their depth profile and energy properties are beyond the scope of this work and may be subject of future research projects.

6.2.2. The n-MOS FET DC Performance

The same experiment has been repeated on an equivalent n-FET structure. As expected, the n-FET is more susceptible to channel ii and shows a small kink in the output characteristic (Figure 6-10) at $t_{Si} = 0.48$ µm. Similar to the p-FET, the n-FET also remains fully functional until thickness values clearly below STI height.

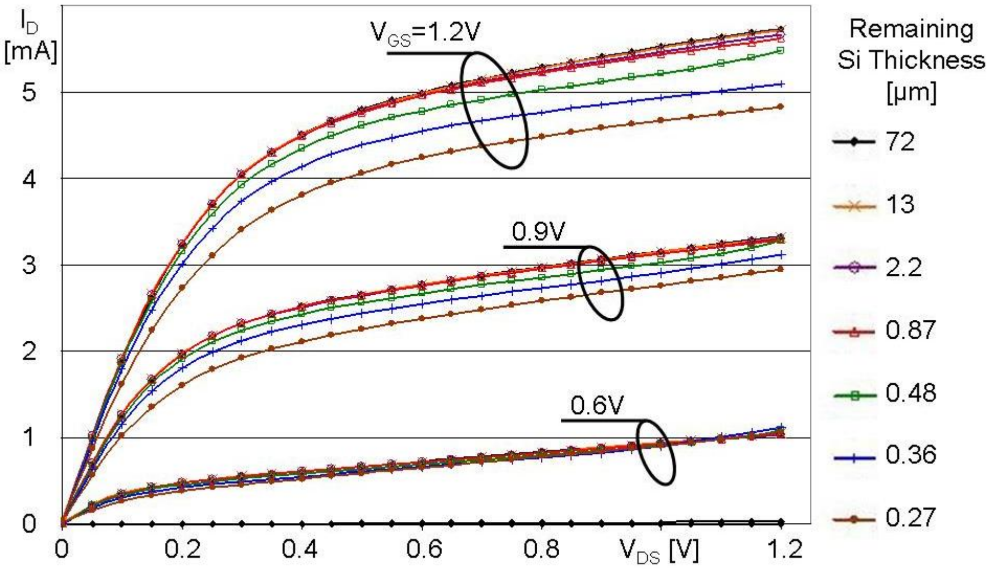

Figure 6-10: Output-characteristic of a 10 µm x 120 nm n-FET, as a function of t_{Si}

Overall, the n-FET follows an equivalent degradation process as discussed for the p-FET. All DC-characteristics remain almost unchanged until t_{Si} approaches a thickness below 2 µm. The linearly extrapolated threshold voltage (plotted together with ΔI_D in Figure 6-11) shows only very small responses to the FIB thinning and increases by less than 15 mV at STI thickness. In contrast to that, the maximum current (ΔI_D, at $V_{GS} = V_{DS} = 1.2$ V) decreases much more and is reduced by ≈ 12% at the same data-point. Especially remarkable is the big difference between the ΔI_D progressions from $V_{GS} = 1.2$ to 0.9 V. As discussed for the p-FET, the slightly stronger degradation at $V_{GS} = 1.2$ V until ≈ 600 nm can be attributed to increased self-heating due to the higher on-current. Below 600 nm, the measured values obey the initial trend (pink dashed line) and the current is increased due to body potential modulation, resulting from ii in the n-FET channel and the increased R_{well}. Since ii is stronger the more a FET is driven in saturation, the I_D increase is higher with $V_{GS} = 0.9$ V compared to 1.2 V. A more quantitative evaluation of the self-heating and the increase of the body potential will follow based on simulations, after the discussion of the diode characteristics.

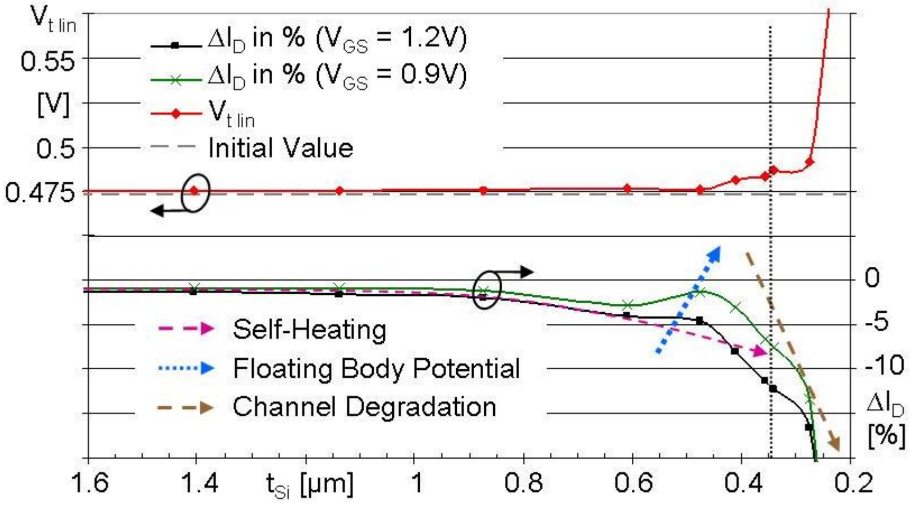

Figure 6-11: I_{Dmax} and $V_{t\,lin}$ as a function of silicon thickness, compared to initial values at $t_{Si} = 72$ µm

Similar to the p-FET measurements, the n-FET diode curves in Figure 6-12 show the same trends, attributed to the increase of R_{well} and the degradation of the junction area.

Following the same strategy as for the p-FET, R_{well} was extracted for the n-FET and is plotted in the upper part of Figure 6-13. In contrary to the p-FET, now the whole p-doped bulk is part of the well, leading to a lower initial resistance value of ≈ 100 Ω. Consequently, R_{well} starts to increase directly with the bulk removal measuring ≈ 156 Ω at 13.2 µm and ≈ 203 Ω at 1.4 µm. Finally, removing the higher doped well also drastically increases the resistance > 1 GΩ with $t_{Si} < 350$ nm. Similar to the p-FET results, R_{well} is proportional to $\exp(1/(t_{Si}))$, as expected due to the retrograded well-doping.

The initial diode reverse current of ≈ $2{\cdot}10^{-13}$ A (at $V_{DB} = 1.5$ V) is again clearly below the in-situ measurement limit. The reverse current increases to more than 1 nA with $t_{Si} = 480$ nm and is finally suppressed by R_{well}. Also these results indicate that the FIB induced imperfections start decreasing the reverse biased diode parameters with a remaining distance of more than 200 nm between the rear surface and the end of the depletion region.

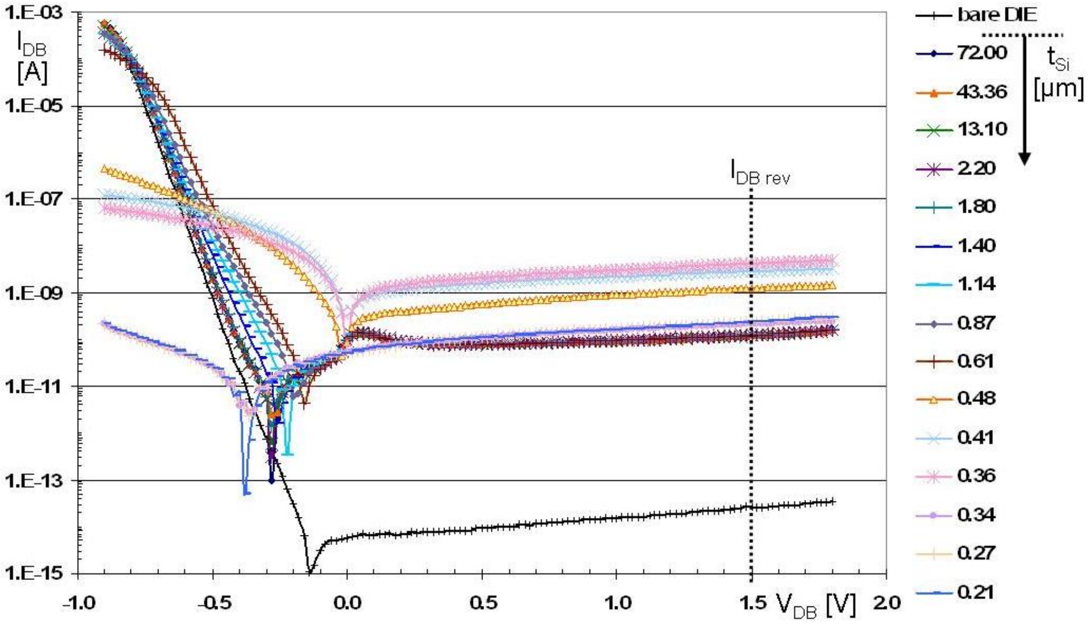

Figure 6-12: Drain to well diode characteristics as a function of silicon thickness, including a reference curve measured on a bare DIE

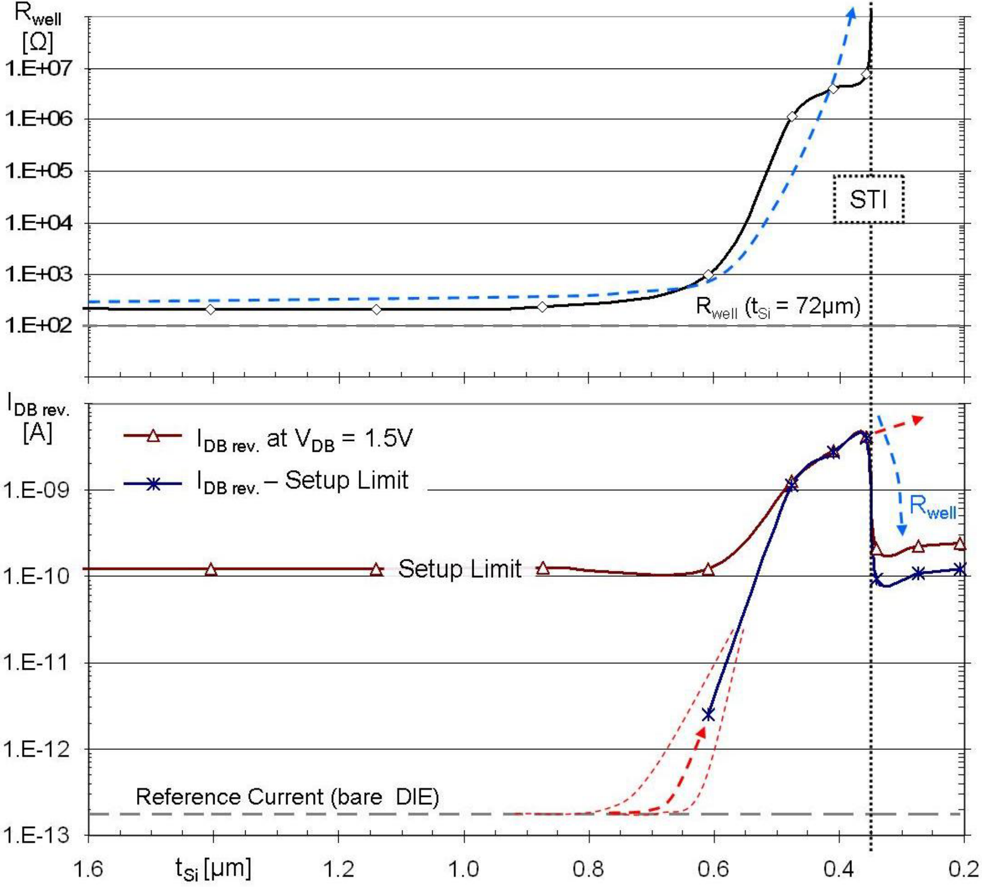

Figure 6-13: Calculated well-resistance (upper plot) and diode reverse current (lower plot) as a function of Silicon thickness

6.3. Simulation Model vs. Measurement Results

The earlier derived DC reference FET simulation models (5.2) will now be used to reproduce the experimental data of the FIB thinning experiment. To underline the degradation theory, the different expected physical impact factors (self-heating, increased well-resistance and diode degradation) will be treated subsequently to allow a quantitative discussion of the various influences.

The first degradation effect is the increased self-heating associated with the subsequent bulk/heat sink removal. With bulk devices, most standard simulation models neglect the heat flux through the backend, reducing the thermal boundary conditions to only one thermal contact at the lower bulk end. But regarding the FIB thinning, combined with all measurements being done in vacuum, the heat flux through the rear surface is almost zero and will not be taken into account. The implemented thermal boundary conditions are depicted in Figure 6-14. All electrical terminals (gate, source, drain and 2 x well) are also assigned to be thermal contacts, having separate heat flux resistance values (R_{th}). R_{th} is analog to an ohmic resistance with the heat flux passing through it being equivalent to an electric current and the resulting temperature difference being equivalent to a voltage drop. R_{th} is defined as l/κ_{th}, with l being the length of the heat conducting material and κ_{th} being its thermal conductivity. Since κ_{th} of SiO_2 is approximately 100 times smaller compared to crystalline Si, the heat flux through the STI was neglected. We have no access to the real values for R_{th} at the source, drain and gate, and the precise determination of these properties would have been beyond the scope of this work. Consequently, the thermal boundary conditions were only roughly aligned to fit the experimental data of three distinguished thicknesses of the thinned n-FET. The n-FET was chosen because of its higher power density and therefore more severe self-heating compared to the p-FET.

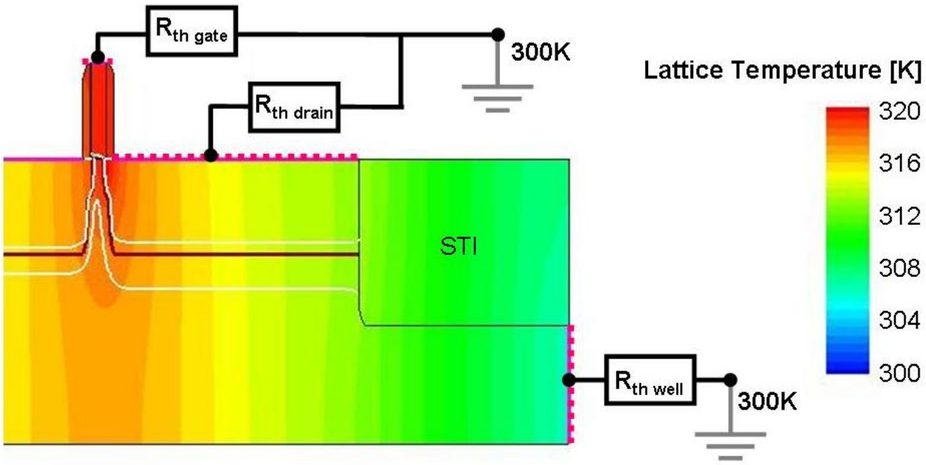

Figure 6-14: FET DC simulation model including the thermal contacts and heat flux resistances

The first alignment point is at the initial FET thickness of 72 µm. According to the thick bulk material, $R_{th\ well}$ was set to a very low value ($5\ 10^{-6}\ cm^2\ K/W$) for the only 3 µm thick simulation reference FET. Based on this initial guess, the other boundary conditions were aligned to fit the measured data. All results are listed in Table 6-1. The second alignment point is at t_{Si} = 610 nm. R_{well} (610 nm) is ≈ 1 kΩ, still much too small to initiate the body charging effects (kink) and also the drain to well junctions are not expected to be affected yet. With respect to the monitored $I_{D\ max}$ values (see Figure 6-11), this measurement point represents the smallest t_{Si} value for which only self-heating can be expected and the simulation results can now be utilized to reproduce and quantify the self-heating effect in the following.

To securely exclude all floating body related effects for this first simulation approach R_{well} was set very low to 10 Ω (per side) and only the FET thickness and thermal boundary condition of the well-contacts were changed. Figure 6-15 shows the simulation results compared to the measured data.

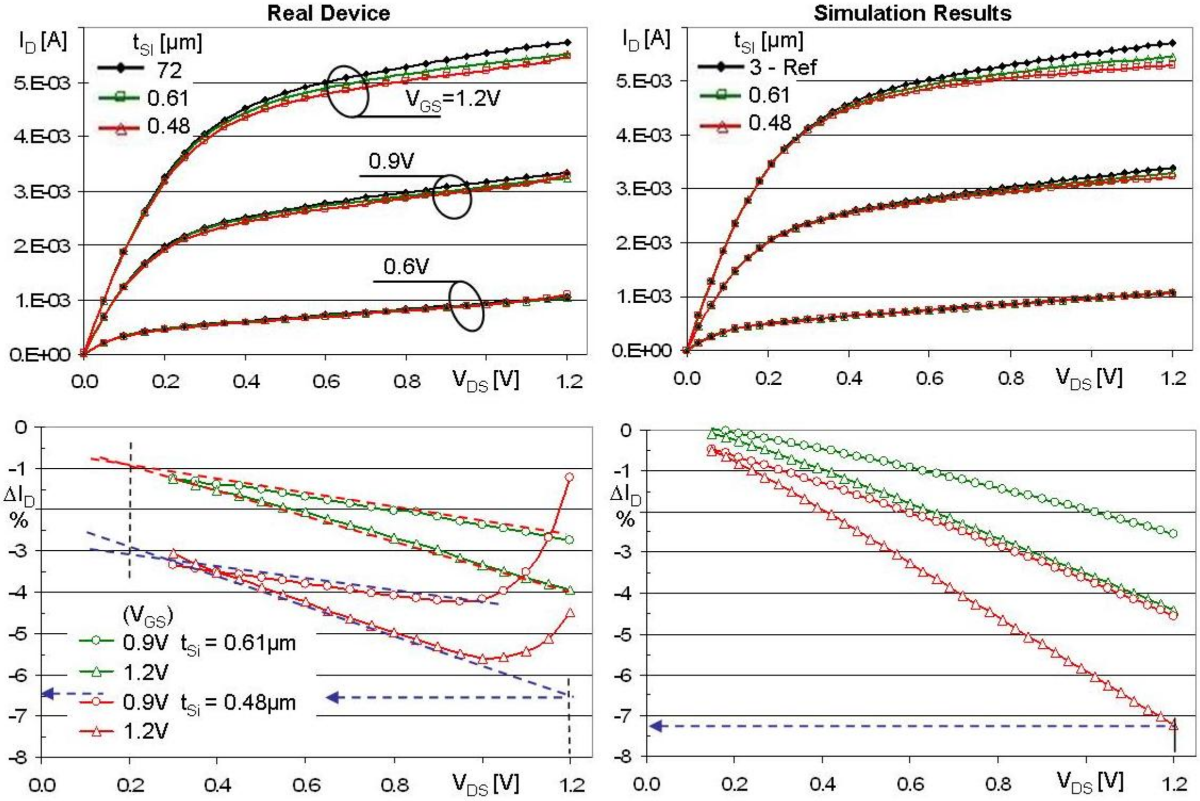

Figure 6-15: Comparison between measured data (left) and simulation results (right). Output characteristics at three different thickness values (upper) and calculated difference between reference performance and thinned devices (lower); $\Delta I_{D(tSi)} = ((I_{D\,(tSi)} - I_{D\,(72\mu m)})*100\% / I_{D\,(72\mu m)})$

The lower diagrams in Figure 6-15 show the normalized difference between the initial drain current (72 µm) and the drain currents measured or simulated at t_{Si} = 610 and 480 nm (in percentage). The data of the real device (left, with t_{Si} = 610 nm) shows a liner performance decrease with V_{DS} in the saturation region of the transistor. Now, the simulation model for the 610 nm FET was adjusted to match the maximum performance decrease, only based on self-heating (no other degradation model included). As expected, the simulation results also show a linear slope for 0.3 V < V_{DS} < 1.2 V but in opposite to the real data, ΔI_D almost drops to zero with V_{DS} < 0.2 V. Extrapolating the linear trend on the real device indicates a degradation offset of ≈ 1% at V_{DS} = 0.2 V. Since self-heating can not be the reason at such low dissipated power (with V_{DS} < 0.2 V) this suggests that the FIB thinning already caused a channel resistance increase of ≈ 1% for t_{Si} = 610 nm. The additional 3% can clearly be attributed to the increased channel temperature.

The degradation measured at t_{Si} = 480 nm (also shown in Figure 6-15) is already more complex. The strong increase of the current for V_{DS} > 1 V can be linked to R_{well} and will be discussed with separate simulations later.

The pure thermal effect can be seen based on the linear extrapolation applied to the V_{GS} = 1.2 V curve (dashed blue lines in the lower left part of Figure 6-15) and indicates a maximum current decrease of ≈ 6.5%, which was also reproduced via simulation with a further increased $R_{th\ well}$. Additionally, the linear approximation towards V_{DS} = 0.2 V indicates that the non-thermal degradation (FIB damage) already increased the channel resistance by more than 2.7% at that data point. Consequently, the simulations for t_{Si} = 480 nm with a pure thermal maximum degradation of > 7% clearly overestimate the self-heating and can be used as a worst case scenario.

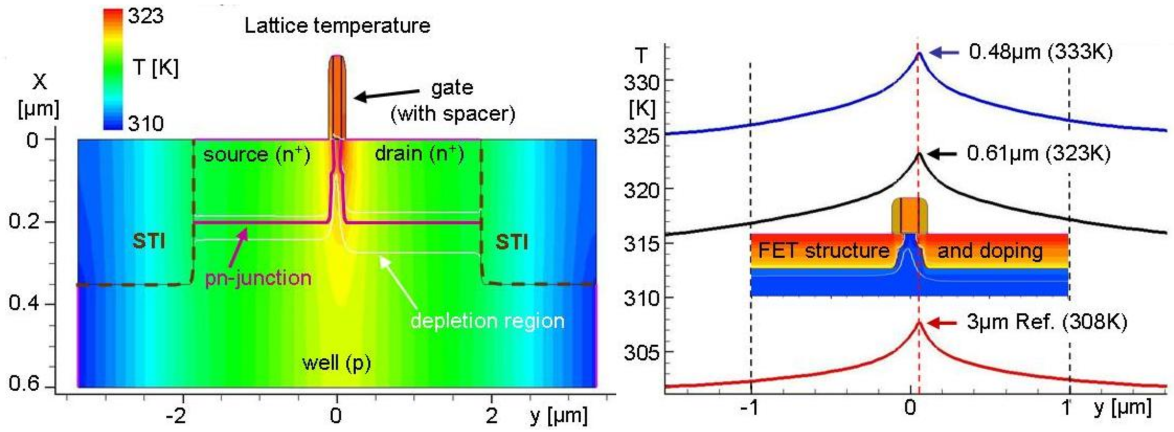

Figure 6-16: Simulated temperature distribution for t_{Si} = 610 nm (left) and line scans (1 nm underneath the gate Si/SiO$_2$ interface) with maximum channel temperature for three thickness values (right) - the line scans are plotted to the same y-scale as the FET structure reflecting the underlying geometry and doping in the right part

The temperature distribution of the simulated n-FET (t_{Si} = 610 nm) and line scans through the three different FET channels are plotted in Figure 6-16. Compared to the reference FET (3 μm), the maximum channel temperature is elevated by 15°C for t_{Si} = 610 nm and by 25°C for the worst case scenario at t_{Si} = 480 nm. As expected, the simulated temperature maximum is found at the drain end of the FET channel.

The geometrical properties of the real test structures during the experiments have been such that the FIB thinning area was always > 3 μm wider than the active area of the FET in x- and y-direction. Assuming t_{Si} to be 500 nm (and subtracting 350 nm for the STI), the aspect ratio of the remaining heat-conducting material would already be greater than 3 μm/150 nm = 60. Consequently, it can be assumed that the self-heating saturates with t_{Si} < 500 nm and remains clearly below the simulation curve for t_{Si} = 480 nm with T_{max} = 333 K.

Table 6-1: Summary of thermal resistance values in accordance to the Si thickness

t_{Si} [μm]	$R_{th\ S/D}$ [cm^2 K/W]	$R_{th\ G}$ [cm^2 K/W]	$R_{th\ well}$ [cm^2 K/W]
72	$1.6 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-6}$
0.61	"	"	$2 \cdot 10^{-4}$
0.48	"	"	$2 \cdot 10^{-3}$

In the next step, R_{well} is increased to confirm the theory that the measured kink is linked to R_{well} in combination with channel impact ionization (ii). R_{well} was chosen in accordance to the extracted values shown in Figure 6-13. With the FET being in saturation, the additional charge generated by ii leads to an increased body potential, reducing V_t, hence increasing I_D. The simulation results for three different R_{well} values of relevant magnitude are plotted in Figure 6-17 with the output-characteristics clearly showing the SOI related kink for $V_{DS} > 1$ V. Comparing the ΔI_D plots to the real measurement results (lower left part of Figure 6-15) shows very good qualitative agreement and a strong dependence on R_{well}.

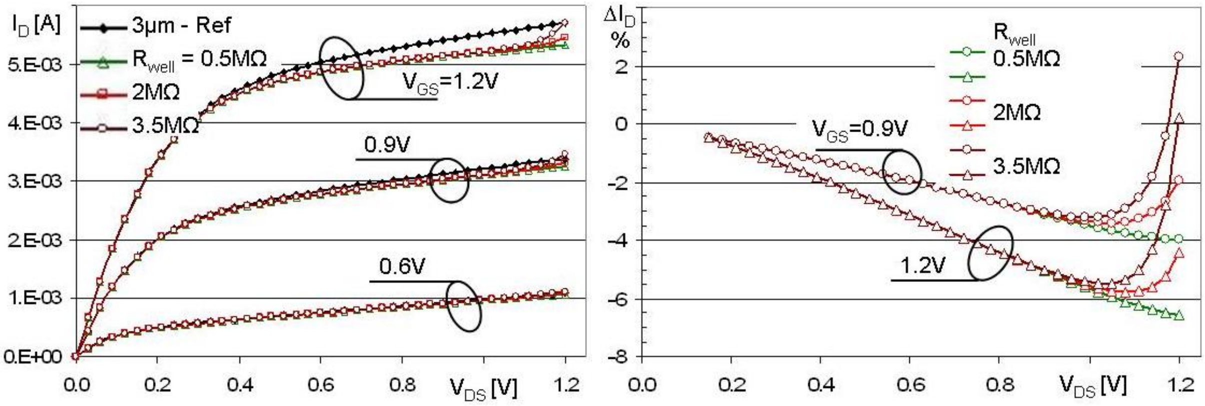

Figure 6-17: Simulation results with increased well-resistance; output-characteristics (left) and normalized difference between reference performance and the 480nm thinned transistor (right)

Due to the necessary simulation model adjustments, discussed in 6-2, ii and other related high field effects may not be computed correctly. The simulations reproduce the measured kink of ≈ 2% with V_{GS} = 1.2 V and ≈ 3% with V_{GS} = 0.9 V in the well-resistance range of 2 MΩ < R_{well} << 3.5 MΩ. Considering the geometrical differences between the real and the simulated structures, the mismatch between measured R_{well} (≈ 1 MΩ) and simulated R_{well} is acceptable and confirms the high accuracy of the underlying simulation model.

The next two data points of the real FET were captured at t_{Si} = 410 and 360 nm. The average thickness of 360 nm marks the preparation step where the STI first became visible, being exposed in ≈ 30% of the FET surroundings. The proposed UtS formation procedure ends here, with the first STI exposure.

As already discussed for the 480 nm thickness, the overall performance decrease is not exclusively due to increased self-heating. For the UtS thickness (t_{Si} = 360 nm), the results plotted in Figure 6-18 show a temperature related current reduction below 6% whereby the increased channel resistance (non-thermal effect) already causes ≥ 7.5%. The almost constant linear threshold voltage (increase of $V_{t\ lin}$ was less than 15 mV as shown in Figure 6-11) implies that the non-thermal degradation is not due to back-bias like effects, induced by charges or Fermi-level pinning on the rear surface, it rather indicates that the thinning reduces the carrier mobility in the channel region (μ_{ch}). This also explains the lower kink with t_{Si} = 360 nm, since the probability for ii scales with the maximum carrier velocity, which is reduced together with μ_{ch}.

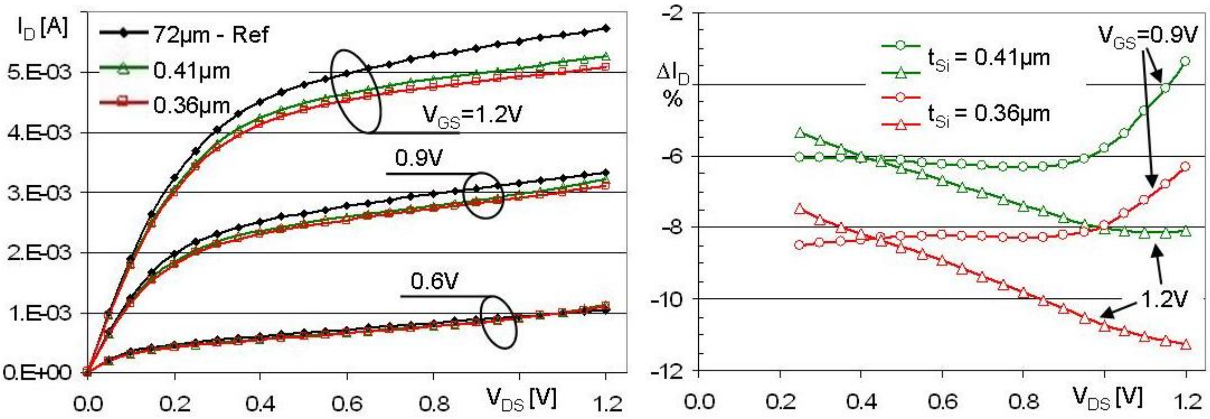

Figure 6-18: Output-characteristics of the real FET and normalized I_D reduction as a function of t_{Si}

Figure 6-19 shows the non-thermal degradation ($\Delta I_{D,nT}$) as a function of the remaining Si thickness, and these results also support the theory that the bulk removal down to ≈ 1 µm does not degrade the intrinsic device parameters, it only affects the thermal boundary conditions, leading to increased self-heating.

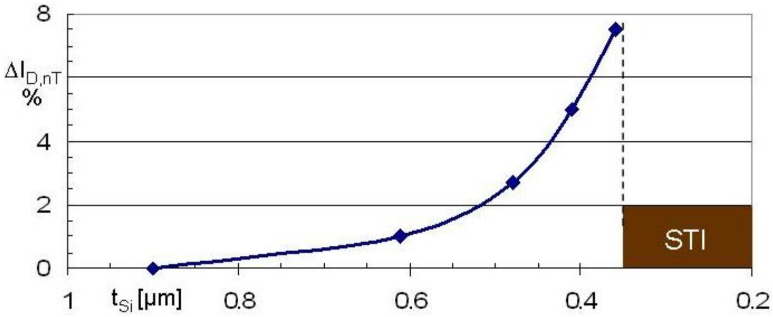

Figure 6-19: Non-thermal degradation for $V_{GS} = 1.2$ V as a function of t_{Si}

The last identified degradation mechanism is associated with the source/drain junctions, can not be measured directly, but its influence can be discussed based on the enlarged view of the normalized current decrease plotted in Figure 6-20.

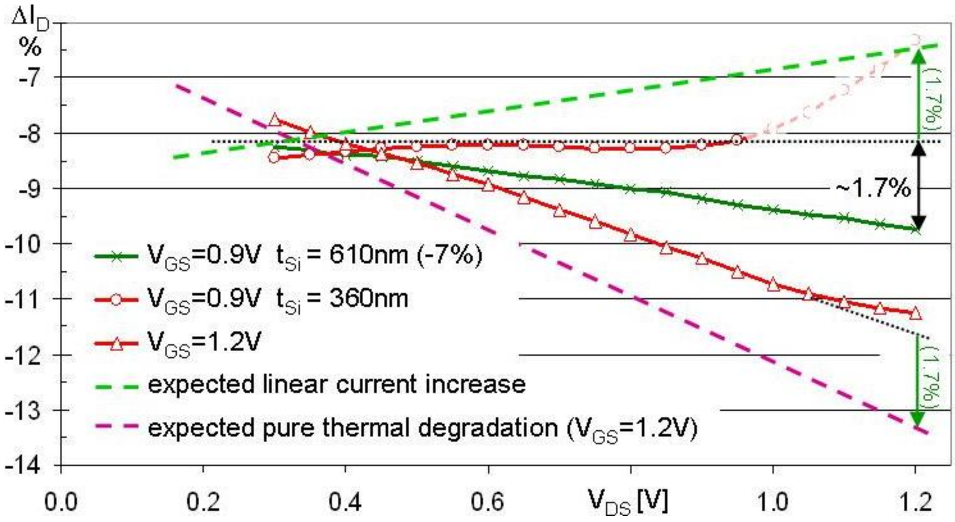

Figure 6-20: Enlarged ΔI_D for $t_{Si} = 360$ nm and parallel shifted curve (-7%) of ΔI_D with $t_{Si} = 610$ nm

First, the pure thermal current reduction, as found earlier for $t_{Si} = 610$ nm ($V_{GS} = 0.9$ V), is used and shifted by -7% to allow for a better overlay to the measured data at UtS thickness ($t_{Si} = 360$ nm). The thermal boundary conditions are worse for $t_{Si} = 360$ nm compared to $t_{Si} = 610$ nm, but regarding the reduced on-current with $t_{Si} = 360$ nm it can be assumed that the pure thermal performance reduction is similar as the reduction measured at $t_{Si} = 610$ nm. With $V_{GS} = 0.9$ V, this ΔI_D can be measured to approximately -1.7%, as indicated by the black arrow in Figure 6-20. This linear current reduction is absent in the $V_{GS} = 0.9$ V branch for $t_{Si} = 360$ nm, which is almost flat until the ii related kink (with $V_{DS} > 1$ V, which shell not be taken into account here). Consequently, there must be an additional effect, equalizing the self-heating related current reduction, causing an almost liner current increase with the slope being similar to the dashed green line. The same effect also reduces the current decrease for the $V_{GS} = 1.2$ V branch. Assuming the same maximum positive influence of +1.7%, the pure thermal degradation would cause a similar current reduction as illustrated by the purple dashed line. Also this increased temperature related maximum on-current drop remains below 7%, not exceeding the simulated worst case scenario with -7.3% and 25°C channel temperature increase.

The discussed current increase is only present for $t_{Si} < 500$ nm. In this thickness range, the increased drain to well diode leakage current $I_{DB\ rev}$ already exceeds 1 nA (see Figure 6-13). In combination with $R_{well} > 1$ MΩ the body potential φ_b will increase significantly. The situation is illustrated in Figure 6-21.

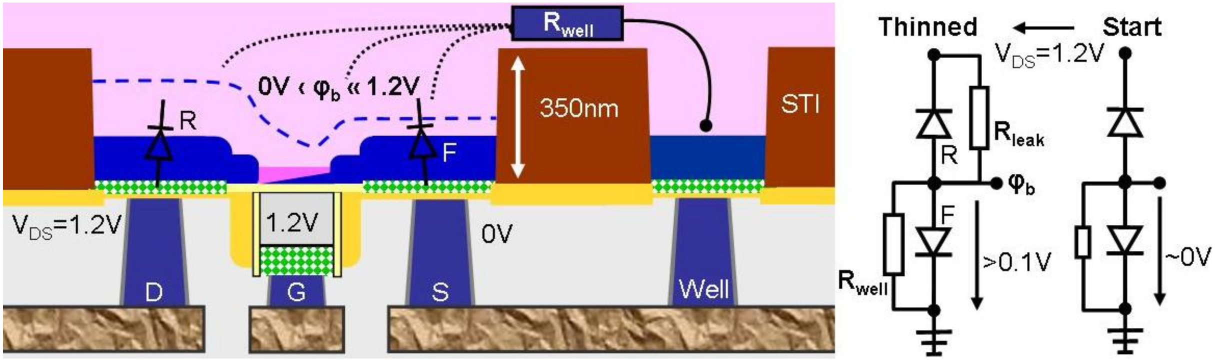

Figure 6-21: Cross-section of a p-FET in static on-state, thinned to little more than STI height

Compared to the initial geometry where the low ohmic well-contact grounds the FET body, the body of the thinned FET is now untied and the static body potential results from a balance between $I_{DB\ rev.}$ (represented by R_{leak} in the schematic), the channel ii related current (not part of the schematic) and the summation of the currents through R_{well} plus the forward current through the also strongly degraded source to well diode. Since the accuracy of the extracted R_{well} values and the measured diode leakage current is rather low in this thickness range, φ_b can not be simply calculated. One way to gain a rough measure of the resulting static body potential of the real test-structures being thinned to STI level is by looking at the characterized body bias sensitivity of an unmodified reference FET (Figure 6-22).

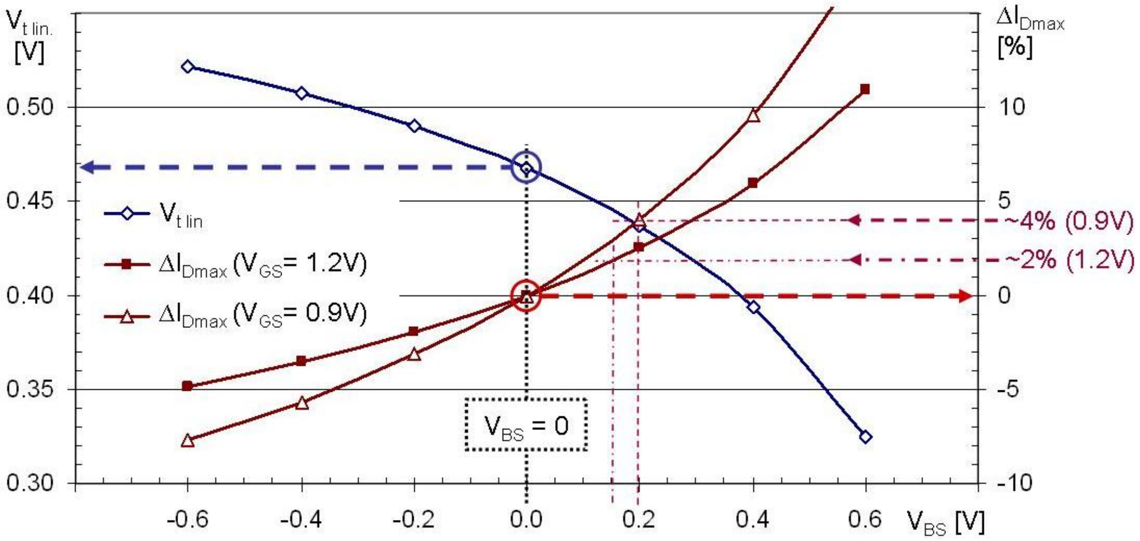

Figure 6-22: Progression of $V_{t\,lin}$ and ΔI_{Dmax} (I_D with V_{DS} = 1.2 V and V_{GS} = 0.9 and 1.2 V) drawn as a function of V_{BS}, measured on a reference/not FIB processed n-FET with identical dimensions (120 nm x 10 µm)

The FIB thinned n-FET shows a body potential related current increase of ≈ 4% for V_{GS} = 0.9 V and ≈ 2% with V_{GS} = 1.2 V (on UtS level, including the kink and the diode degradation effect, see Figure 6-20). Using these values in the plot of the measured back-bias sensitivity of the reference n-FET delivers the floating body potential to be increased by ≈ 200 mV and 150 mV for V_{GS} = 0.9 V and 1.2 V respectively.

The meaningful simulation of the ion beam based degradation requires the precise knowledge about the nature of the FIB induced imperfections, e.g. the concentrations and profiles of crystalline point- or line defects, traps, Gallium contamination and radiation damage and their energy levels in the Si bandgap. The extraction and implementation of all necessary parameters is clearly beyond the scope of this work and may be addressed in following projects.

6.4. Summary of DC Performance Degradation

The combination of experimental results with physical device simulations allowed identifying several different effects which alternate the FET device physics under DC operation. The bulk removal leads to increased self-heating with a simulated maximum temperature increase of 25°C for the n-FET channel in static on-state, compared to the unmodified FET. This results in a temperature related maximum current reduction of ≤ 7% on the final UtS level.

Additionally, the well-contact resistance increases as a function of the remaining Si thickness, entering the GΩ regime with $t_{SI} < 0.5$ µm. The increased resistance initiates a kink in the output characteristic of the n-FET causing an intermediate positive effect of up to 3% in the small thickness range 0.5 µm > t_{SI} > 0.4 µm. The p-FET is less susceptible to ii in the pinched-off channel region and does not show this effect.

The FET channel mobility starts decreasing with $t_{SI} < 1$ µm due to the FIB induced imperfections which caused a current reduction of ≤ 8% on UtS level. This mobility reduction also explains the reduced kink effect during the final preparation phase (t_{SI} > 0.4 µm).

Finally, the FIB induced leakage of the drain/source to well diode junctions elevates the body potential, reducing V_t, hence increasing the maximum current by $\leq 2\%$ on the final UtS level. The effect is similar to the ii related kink but not as V_{DS} dependent and present on both FET types.

This results in a maximum current reduction of $\approx 12\%$ and $\approx 7\%$ for the n-and p-FET respectively, which may already be problematic considering analog devices, but for most circuit edit and analysis application in CMOS environments this is rather uncritical.

7. Influence of Backside FIB thinning on Dynamic Circuitry Performance

The most commonly used test structures for the characterization of dynamic circuitry performance are ring oscillators. In the following, experiments on three differently designed ROs will be used to evaluate the impact of the proposed FIB thinning procedure on high frequency circuitry operation. Combining experimental results with dynamic device simulations will allow deriving a theoretical model explaining the measured performance alterations.

7.1. Experimental Setup and Test Structures

Figure 7-1 shows the complete layout of RO (A), including a five stage output buffer, a ten stage divider (1/1024) and the high frequency ring comprised of 64 identical inverters (see Figure 7-3) plus one NAND gate (with enable as second input), in total 65 stages. To allow for a real quantitative measure of the performance modulation, all 65 stages, occupying an area of more than 50 x 25 µm^2, have to be thinned at the same time with the highest possible planarity between the trench bottom and the device layer. The key challenge to gain the necessary planarity in such wide areas is to control the global trench to n-well level. On RO (A), the global trenching was stopped at a remaining thickness of 1.43 µm with an almost perfectly flat center part. The area in which the thinning was continued (red dotted rectangle in the right part of Figure 7-1) had a maximum planarity mismatch of less than 70 nm (< 0.1%, across the diagonal of ≈ 90 µm), proven by the absence of interference patters (see subsection 3.2.2 for details). But even with such level of flatness, all later thickness values have to be understood as being average values, derived by measurements of interpolation points and linear extrapolation as already discussed for the experiments on single FETs in 6.2.

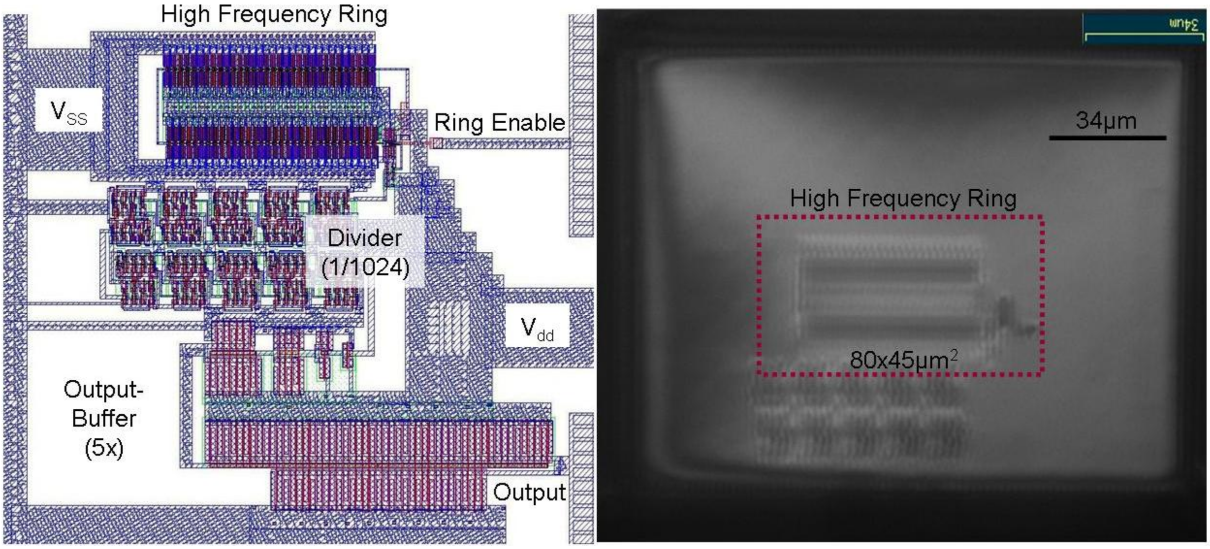

Figure 7-1: Layout of 120 nm technology RO (A) test structure comprised out of 65 inverters (left) and IR-optical image of the same ring (right), captured in the FIB at n-well level (t_{Si} =1.43 µm). The red dotted rectangle indicates the preparation area for the following thinning to STI level.

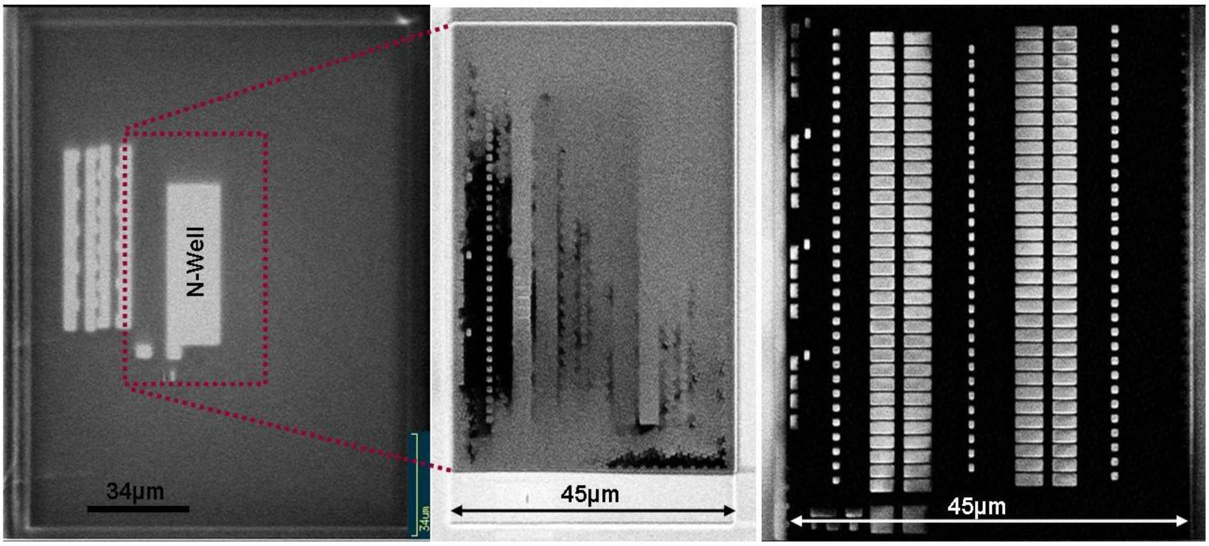

Figure 7-2: FIB images showing the trench to n-well (left) and the smaller opening down to STI level at an average thickness of ≈ 355nm (middle) and ≈ 250nm (right), all captured at RO (A)

Figure 7-2 shows FIB images of RO (A), being rotated by 90° compared to Figure 7-1. The left one was captured on n-well level, also showing the area of further thinning (red rectangle). The middle part was captured after 720 s of etching towards STI level and as can be seen, the surrounding STI is revealed at almost 50% of the ring. Continuing the etching process for 15s more, the whole ring (or surrounding STI) was exposed. Consequently, with the calculated etch-rate being ≤ 1.5 nm/s, the planarity mismatch in the ring area was less than 30 nm at STI height.

7.2. RO (A) with Speed Optimized Inverters

The thinned inverters of RO (A) are fabricated in the speed optimized folded design and the most important layout parameters are marked in Figure 7-3. The equivalent effective drain length (L_d, see Figure 4-8) of a similarly performing non-folded FET would be 360 nm/2 = 180 nm as discussed in 4.2.2.

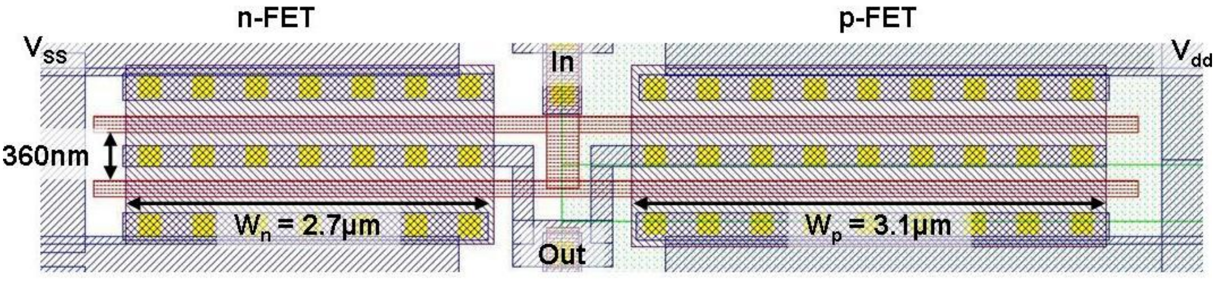

Figure 7-3: Enlarged layout of one out of the 65 inverters of RO (A), with speed-optimized folded design; without additional load capacitances, FO = 1

Throughout the whole FIB procedure, the DuT was powered up by use of vacuum setup A (see appendix A) and the resulting RO frequency was monitored with a conventional Oscilloscope. The frequency was measured for the different supply voltages 0.8, 1.2 and 1.6 V (Figure 7-4), whereby 1.2 V is the nominal supply voltage for this technology node. Similar to the results on single FETs, the thinning process down to n-well level does not change device physics and the ring performance also remains almost unchanged down to a remaining silicon thickness of ≈ 700 nm. Below this thickness, the ring experiences a strong performance increase with a maximum at t_{Si} values from 500 nm down to 400 nm. The sensitivity to the FIB induced delay reduction is clearly V_{dd} dependent with $\Delta f_{Out\ max}$ ≈ 37% for V_{dd} = 0.8 V, 17% for V_{dd} = 1.2 V and only ≈ 11% for V_{dd} = 1.6 V.

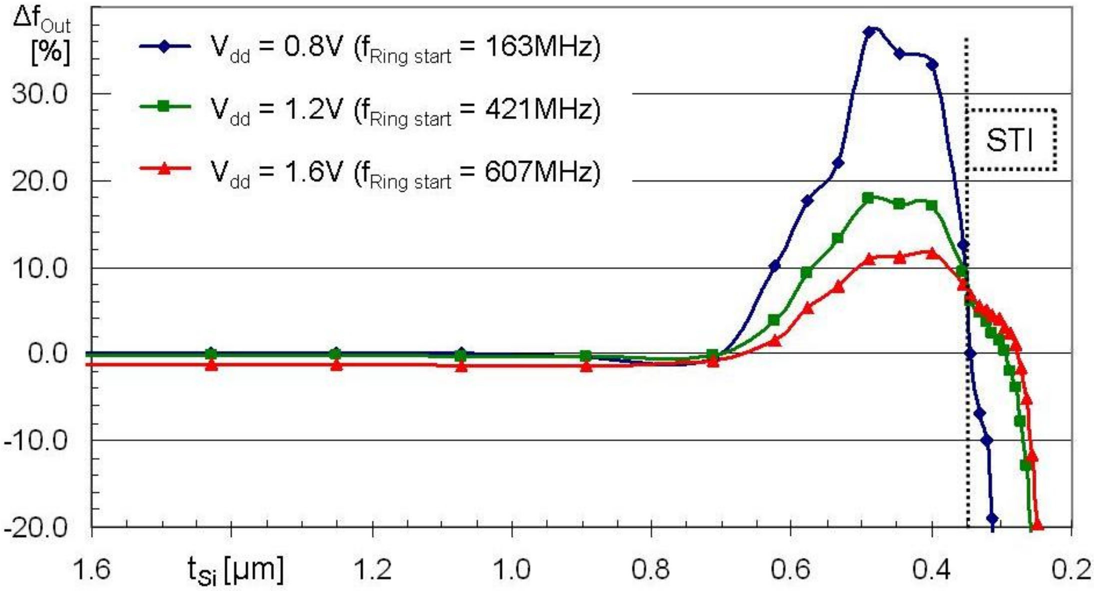

Figure 7-4: Progression of RO (A) frequency, measured at three different supply voltages and drawn as a function of t_{Si}

With all V_{dd} values, the positive effect is strongly reduced to ≈ 10% when t_{Si} approaches STI height and drops below zero later on. Also in this thickness range, the circuitry is more susceptible to degradation with the lower supply voltage.

Comparing these results to the DC performance alteration of the single p- and n-FET discussed earlier shows that the increased circuitry speed can not be explained only based on these static models, since the overall FET performance never exceeded the initial conditions. In the following, dynamic physical device simulations will be used to gain a deeper understanding of the experimental results.

7.3. Simulations based on RO (A)

The goal of the following simulations is to qualitatively reproduce the measured results, based on the adjusted DC transistor models and the known RO details as its geometry and initial performance. The geometry of RO (A) was already depicted in Figure 7-3, having a width ratio between p- and n-FET of 3.1 µm/2.7 µm and an effective drain length L_d = 180 nm.

In contrary to the single FETs, the ROs have a much reduced power dissipation density. This is based on the CMOS architecture in combination with the high number of inverters being part of the ring. With a well designed CMOS gate, the static current is very close to zero, reducing the static power consumption to an irrelevant level. Power is only consumed during switching whereby the ratio between the switching time with relevant self-heating and the idle period is below 1/65 (regarding a full inverter stage – n- and p-FET together). Compared to the DC simulations where the maximum temperature increase remained below 25°C with thinning to STI level, the expected self-heating of the ROs would be clearly below 1°C in average and will be neglected. Consequently, all dynamic simulations are carried out excluding thermal aspects.

Based on the initial RO frequency measurement, the period of oscillation τ_{osc} and the average inverter delay τ can be derived as discussed in 4.3, and shown here only for the results at the nominal supply voltage (V_{dd} = 1.2 V):

$$\tau_{osc} = \frac{1}{421MHz} \approx 2.375ns \tag{7-1}$$

$$\tau = \frac{\tau_{osc}}{2 \cdot n} \approx \frac{2.375ns}{2 \cdot 65} = 18.277\,ps = \frac{(\tau_n + \tau_p)}{2} \tag{7-2}$$

Table 7-1 sums up the known initial RO performance values, based on the measurement results of Figure 7-4. For simplicity, the oscillation period was rounded.

Table 7-1: Calculated oscillation period and average inverter delay

V_{dd} [V]	RO (A) f [MHz]	τ_{osc} [ns]	τ [ps]
0.8	163	≈ 6.14	47.245
1.2	421	≈ 2.4	18.277
1.6	607	≈ 1.27	12.668

As discussed in the introduction of the simulation environment (5.3), only one isolated inverter stage is simulated at a time. In a real RO, the oscillation propagates as a normalized signal, where the transition speed of one inverter output determines the input rise- and fall-time for the following stage. But the RO output signal does not allow a direct measurement of the inner transition times due to the presence of the divider and output buffer. Consequently, the normalized rise and fall times had to be determined by initial simulations. Furthermore, the capacitive load associated with the following stage and the interconnecting metal ($C_{L\ sim} = C_{in} + C_L$) can only be approximated by the gate oxide capacitance of one inverter stage:

$$C_{L\_sim} \approx C_{OX} = \frac{\varepsilon_{SiO2} \cdot L_G \cdot (W_{G\_n} + W_{G\_p})}{t_{OX}} \approx 11\,fF \tag{7-3}$$

(with t_{ox} = 2.2 nm) and needed to be aligned to achieve good agreement with the initial average inverter delay τ.

7.3.1. Alignment of the Reference Inverter Model

An inverter with 3 µm Si thickness served as a reference structure. Different rise and fall times were tested and some results are shown in Figure 7-5 for example. The simulation input is the time period in which the input signal changes linearly from either zero to V_{dd} or vice versa, measured at transition A to be 45 ps. The resulting transition times of the output signal change with the input rise and fall times (t_{rise} and t_{fall}). The transition time was measured between 10% and 90% of the output signal, as shown at transition B. For any given RO, there is only one set of rise- and fall-times representing the normalized oscillation signal. The inverter delay is measured at the 50% level, as depicted at the switching event C.

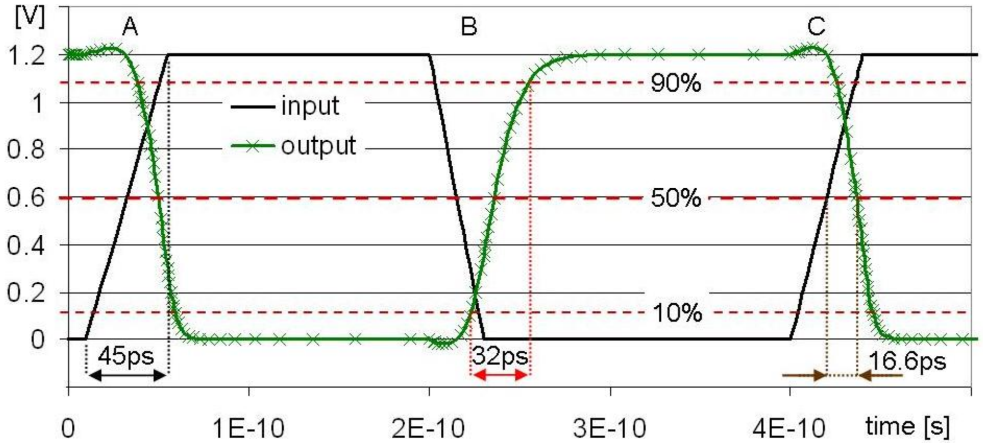

Figure 7-5: Simulation results of 3 µm reference inverter with different rise and fall times; $C_{L\,sim}$ = 13.5 fF

To find the optimum for t_{rise} and t_{fall} in combination with $C_{L\,sim}$, various iterative simulations were done and the final results are illustrated in Figure 7-6. This plot correlates the input transition time with the resulting output transition for the rising and falling edge. Considering a fall transition at the inverter input, a transition time of ≈ 19.5 ps on the y-axis (as marked by the dashed blue line) results in an output rise time of ≈ 30 ps on the x-axis. In case of a rise transition at the input the x- and y-axis have to be read in the opposite order, following the red dashed lines.

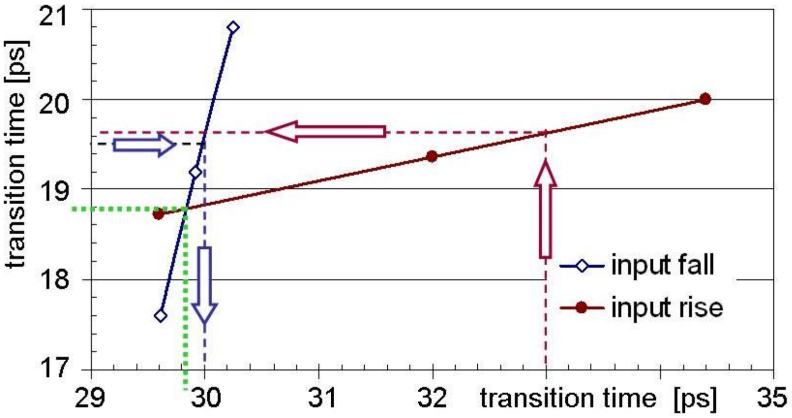

Figure 7-6: Correlation between the 10% to 90% transient times for three different falling and rising transitions, allowing extracting the system inherent normalized rise- and fall-times

Based on these results, the best approximation of the normalized signal is found at the intersection. Due to the definition of the transition time (measured between 10% and 90%), the corresponding rise- and fall-times (0 to 100%) can be calculated by dividing with 0.8 to:

$$t_{rise} = \frac{29.8\,ps}{0.8} \approx 37.3\,ps \tag{7-4}$$

$$t_{fall} = \frac{18.8\,ps}{0.8} = 23.5\,ps \tag{7-5}$$

Using these rise and fall times in combination with $C_{L\,sim}$ = 13.5 fF and the RO equivalent oscillation period τ_{osc}= 2.4 ns yields the simulation results shown in Figure 7-7.

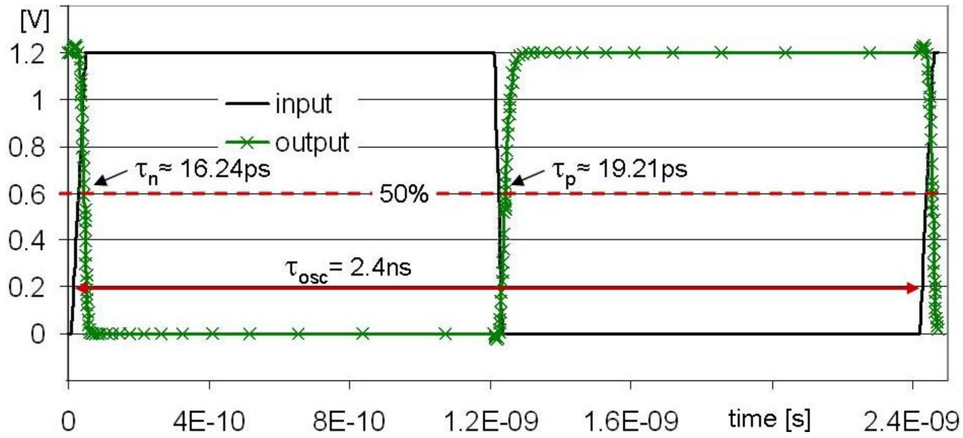

Figure 7-7: Reference inverter, aligned to the initial performance of RO (A)

The average single stage delay $\tau = (\tau_n + \tau_p)/2$ was measured to be ≈ 17.73 ps, less than 3% off compared to the real RO. These results will serve as a reference for the simulation of the thinned structures.

7.3.2. The Thinned Inverter Model

With the real RO (Figure 7-4), the frequency starts increasing with t_{Si} below 700 nm and reaches a maximum at around 450 nm. Also, the extracted well-resistances of the single n- and p-FETs show a similar behavior (see Figure 6-8 & Figure 6-13). Two data points were chosen for the simulation: the first one at t_{Si} = 500 nm and the second one at t_{Si} = 450 nm in the thickness range with the maximum positive effect. Figure 7-8 shows the implemented n- and p-FET, being combined to a CMOS inverter. Despite the well-contacts, all terminal connections and also the load capacitance $C_{L\ sim}$ were kept the same compared to the reference setup. But due to the changed dynamic behavior it becomes much more complex to align the internal signal in accordance to the normalized RO signal propagation. Therefore, before discussing the identified reasons for the speed gain, the initial model alignment has to be explained first.

The simulation tool scales any assigned terminal resistance with the width of the actual device, here 2.7 µm and 3.1 µm for the n- and p-FET respectively. The first simulations for t_{Si} = 500 nm were carried out with the assigned well-resistance of $R_{well\ left} = R_{well\ right}$ = 10 MΩ, resulting in an effective well-resistance of 1.8 MΩ for the n-FET and 1.6 MΩ for the wider p-FET. Because of the fundamentally different geometries compared to the single FETs, these R_{well} values are only rough estimations of the expected well-resistances based on the measurements of the single FETs. The R_{well} dependence will be evaluated later.

The simulation results (Figure 7-8) show the so called "electrostatic potential" captured during the dynamic simulation. The "electrostatic potential" represents the local difference between the intrinsic Fermi-level and the resulting simulated Fermi-level. For example, the source area of the n-FET (left) is on ground potential whereby the Fermi-level is close to the conduction band edge due to the high n-type doping, resulting in a local potential of approximately half the bandgap ≈ 0.55 V. The source of the p-FET would have a potential of approximately $-W_g/2$ because of the high p-doping, but due to the applied supply voltage of V_{dd} = 1.2 V the local potential is increased to ≈ 0.6 V.

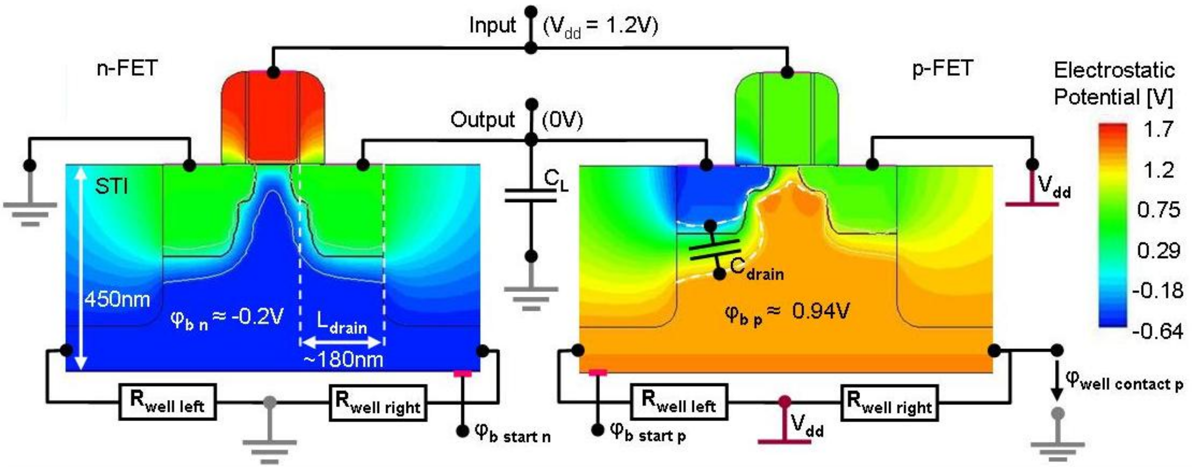

Figure 7-8: Dynamically simulated inverter with FET extrinsic circuitry

Due to this definition, the simulated potential values can be misleading and in the following, especially the body potential φ_b will be discussed according to the inner potential of the high ohmic well-contact $\varphi_{well\ contact}$ as illustrated for the p-FET. Due to the relative low ohmic well-material in combination with the very low currents (< 1µA because of R_{well} >1 MΩ), the inner contact potential is equivalent to the relevant body potential with an error of less than 1 mV and will be used as φ_b throughout all later discussions.

During a simulated switching event, for example from input low to high, the potential of the common drain node has to fall from high to low. During this process the drain to well junction capacitance C_{darin} has to be unloaded in the n-FET and loaded in the p-FET. But due to the high well-resistance values (>1 MΩ), the displacement current associated with the loading and unloading of these junction capacitances causes a voltage drop across R_{well} and consequently, the body potential becomes a strong function of time.

The upper part of Figure 7-9 shows the simulation input and the corresponding output signal, whereby the lower part shows the resulting body potential of the n-FET ($\varphi_{b\,n}$) and the body potential of the p-FET -($\varphi_{b\,p}$ - 1.2 V), which was shifted by -1.2 V to plot it in the same scale as $\varphi_{b\,n}$ and inverted (multiplied by -1) to allow a more intuitive interpretation of the resulting effects on the transistors which will be discussed later on.

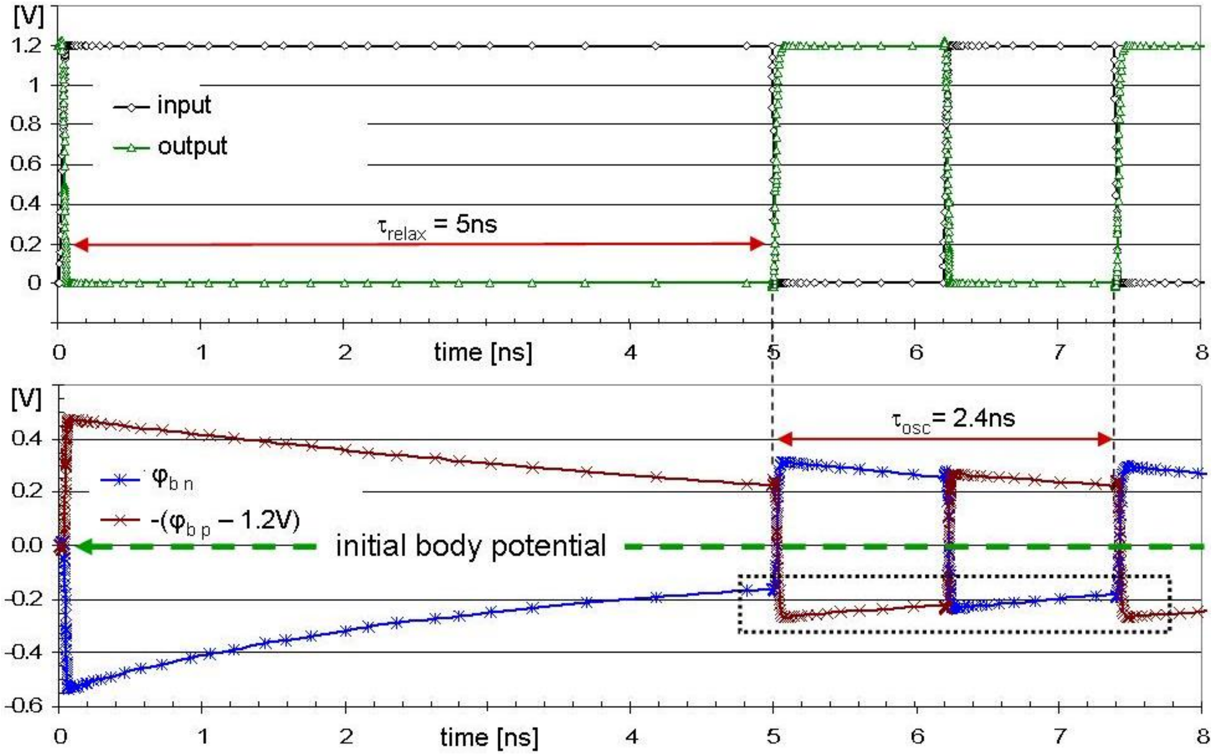

Figure 7-9: Simulation results for RO (A), with $t_{Si} = 500$ nm including input and output signal (upper) and the body potential of the n- and p-FET (lower); $\varphi_{b\,p}$ was shifted by -1.2 V and inverted

In the real ROs, the signal propagates at the RO dependent maximum speed and the average body potential of the n- and p-FET will automatically reach a stable value after some hundred switching events in less than a second after turning on the supply voltage. In contrast to that, the dynamic simulation is limited to the calculation of only a few subsequent switching events due to the long simulation time (>2 h per single transition).

Furthermore, also the dynamic simulation starts with a quasi-static part, similar to the earlier DC FET simulation, where the inverter is ramped up into the start condition of the dynamic operation. Based on the underlying DC FET models, the body potential of the n- and p-FET results in being close to neutral ($\varphi_{b\,n} \approx 0$ and $\varphi_{b\,p} \approx 1.2$ V), as defined by the well-contact potentials. With the first dynamic switching event, the body potential follows the output to some extent and both, $\varphi_{b\,n}$ and $\varphi_{b\,p}$ are excited by more than 0.4 V, as can be seen in the lower part of Figure 7-9. In the following idle period, the body potential slowly relaxes. Waiting for a certain time τ_{relax} until the next transition is performed allows to approach the system inherent average body potential much faster than by continuously switching with the given ring oscillation period τ_{osc} right up from the beginning.

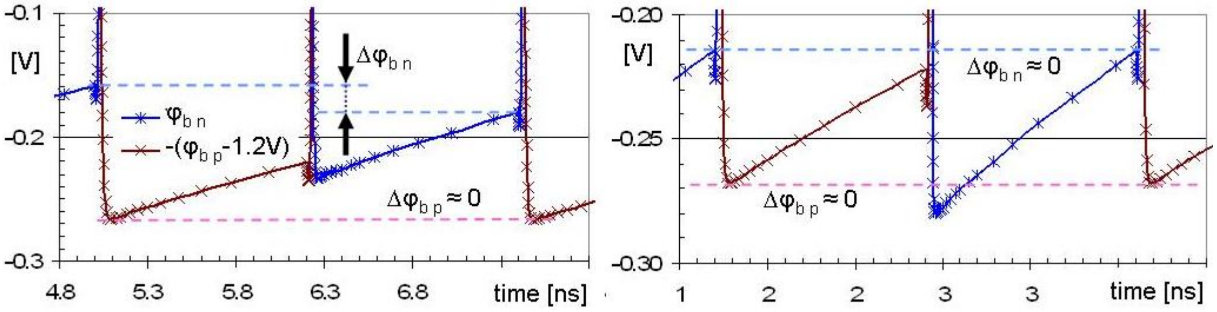

Figure 7-10: Enlarged view of the body potential, used for model alignment by minimizing $\Delta\varphi_{\mathrm{bn}\ \&\ \mathrm{p}}$

The enlarged view in the left half of Figure 7-10 shows the result of this approach. The initial relaxation time of τ_{relax} = 5 ns is sufficient to bring $\varphi_{\mathrm{b\,p}}$ into the desired condition where there is almost no difference between the two extreme $\varphi_{\mathrm{b\,p}}$ values of two subsequent switching cycles, hence the p-FET has reached a stable average body potential and $\Delta\varphi_{\mathrm{b\,p}} \approx 0$. Unfortunately, only the body potential of one FET, either n- or p-type, can be aligned at a time based on this technique ($\Delta\varphi_{\mathrm{b\,n}} > 20$ mV) and the relatively slow relaxation during the subsequent switching at τ_{osc} would enforce the simulation of >10 cycles before the average value of $\varphi_{\mathrm{b\,n}}$ also reaches acceptable stability.

To allow for faster and more accurate simulations, the two additional contacts $\varphi_{\mathrm{b\,start\,n}}$ and $\varphi_{\mathrm{b\,start\,p}}$ were added to the inverter model, already depicted in Figure 7-8. By use of these contacts, the two FET body potentials can be ramped to any desired start value before the dynamic part of the simulation begins, allowing aligning $\varphi_{\mathrm{b\,n}}$ and $\varphi_{\mathrm{b\,p}}$ separately. During the dynamic part, these contacts are floating and do not interfere with the body potential anymore. The achieved results for the 500 nm thin inverter are shown in the right part of Figure 7-10 where the mismatch was measured to be less than 0.5 mV for both FETs.

Since the thinned structures show an increased speed compared to the reference device, the input rise and fall times also had to be adjusted to match the normalized signal conditions as described for the reference inverter, but here already based on the aligned average body potentials.

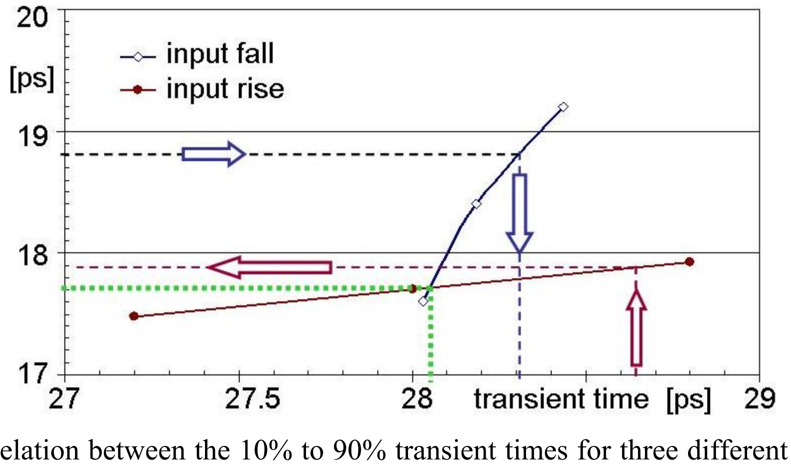

Figure 7-11: Correlation between the 10% to 90% transient times for three different falling and rising transitions, simulated for the 500 nm thin inverter

$$t_{rise} = \frac{28.05\,ps}{0.8} \approx 35\,ps \tag{7-6}$$

$$t_{fall} = \frac{17.7\,ps}{0.8} \approx 22\,ps \tag{7-7}$$

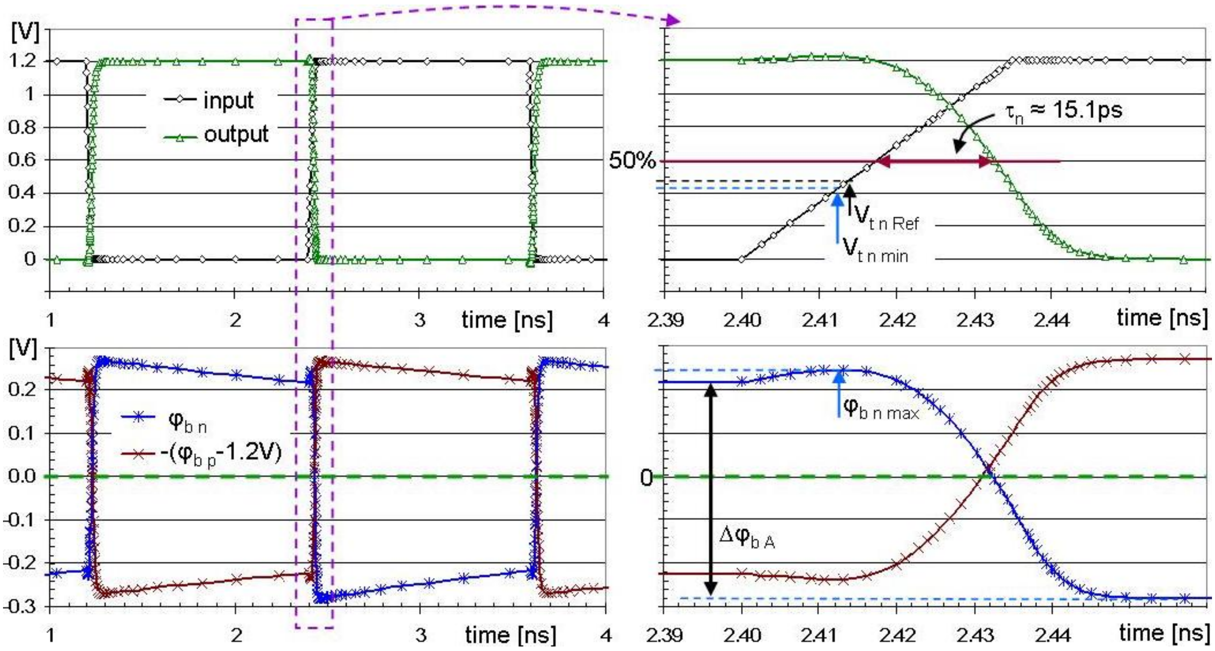

Figure 7-12: Simulation results for t_{Si} = 500nm

The results presented in Figure 7-12 were simulated with respect to all former model alignments and show a delay of $\tau_n \approx$ 15.1 ps and $\tau_p \approx$ 17.46 ps. Compared to the reference inverter this marks a delay reduction of approximately 8.2%, showing the same tendency of delay reduction as the measurements but not yet the right amplitude.

7.3.3. Identified reasons for the inverter performance increase

The simulated speed gain can be explained based on the modulated body potential and will be discussed focusing on the n-FET dominated falling edge (of the output signal), plotted in the right half of Figure 7-12. During the transition, the n-FET has to discharge the common drain node, including the load capacitance $C_{L\,sim}$ and both drain capacitances $C_{drain\,n}$ and $C_{drain\,p}$. As discussed in the theoretical part (4.2.1), the switching performance is influenced by the threshold voltage and the maximum current, both being affected by the body potential. The two simulated body potentials are plotted in a way that a resulting value above zero marks a positive influence on the FET performance, hence for the n-FET V_t is reduced and I_{Dmax} increased until approximately half of the output transition is performed. The p-FET experiences the exact opposite. With the shown falling edge, $\varphi_{b\,n}$ reaches its maximum value of $\varphi_{b\,n\,max} \approx$ 0.24 V right in the short time period where the n-FET is turned on. Based on the body-bias sensitivity measured for a single FET (see Figure 6-22), the threshold voltage is reduced by more than 40 mV compared to the reference value. With the linearly rising input signal (t_{rise} = 35 ps), the input voltage surpasses V_t about 1.166 ps earlier because of the 40 mV V_t reduction. The effect is illustrated in the upper part of the image. Consequently, the whole switching process starts earlier. Furthermore, the n-FET performance remains increased compared to the reference device as long as $\varphi_{b\,n}$ is above zero, so almost until the output voltage reaches V_{dd} / 2, resulting in faster discharging of the output node due to the increased maximum current. But during the second half of pulling down the output voltage, $\varphi_{b\,n}$ drops below zero and the n-FET has a worse performance compared to the reference device. Since the transition of the following stage is already initiated, the reduced FET on-current has less negative effect on the ring performance compared to the initial V_t reduction and the overall effect remains positive. A good measure for the ratio between positive and negative effect is the level of the average body potential which results to be slightly positive for the simulated data with $V_{dd} = 1.2$ V, shown in Figure 7-12. The average body potential is V_{dd} dependent and additionally affected by the diode degradation and will be reviewed once more later on.

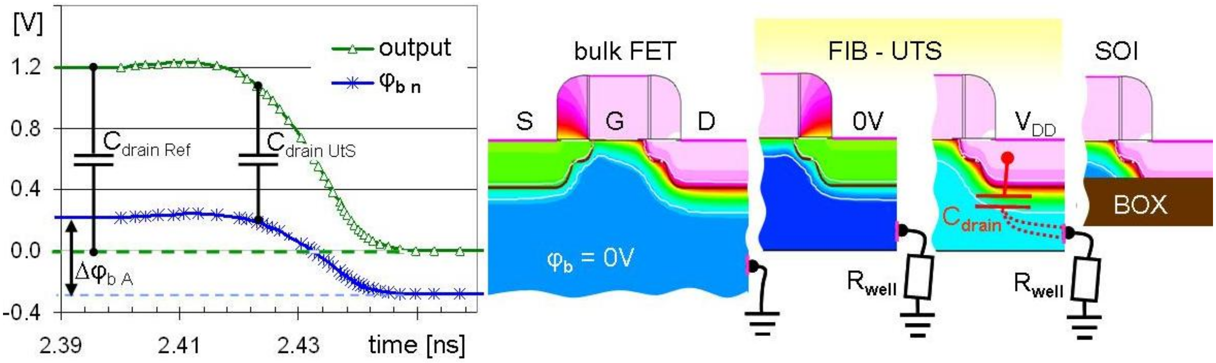

Figure 7-13: Simulated output signal and corresponding body potential of the 500 nm n-FET (left) and illustrations of one bulk- two UtS- and one SOI-FET (right)

The second reason for the performance increase can be linked to the drain to well junction capacitances $C_{drain\,(n\,\&\,p)}$ which build approximately 20% of the output load (calculated based on simulation results, which are discussed later, see Figure 7-16). Figure 7-13 shows the simulated output signal in direct comparison to the body potential of the 500 nm thin n-FET during the falling edge transition. With an unmodified bulk FET, φ_b is approximately zero regardless of the transistor operation. Consequently, when the output voltage is pulled down from $V_{dd} = 1.2$ V to zero, also C_{drain} has to be fully discharged with $\Delta V_{Cdrain} = 1.2$ V. With the FIB thinned transistors the body potential follows the output voltage in a way that ΔV_{Cdrain} is reduced by $\Delta\varphi_{b\,A}$ during the transition.

Real SOI devices, as depicted on the right of Figure 7-13, show an approximately 20% higher performance compared to bulk technology based on the full removal of the lower drain junction as already discussed in 4.4.

With $\Delta\varphi_{b\,A} \approx 0.52$ V in this example, ΔV_{Cdrain} is reduced by $\approx$ 43%, which will be close to but must not be equivalent to the reduction of the necessary displacement current because of the voltage dependence of C_{drain} itself.

The pure displacement current resulting from discharging C_{drain} is not directly accessible since it is not an isolated terminal current but one out of several contributions. Following the simplified model depicted in Figure 4-8, with all terminal currents being positive when flowing into the device, the drain and source terminal currents are:

$$I_S \approx -I_{ON} + I_{Csource} - I_{ch\_S} - I_{Cov\_S} \tag{7-8}$$

$$I_D \approx I_{ON} + I_{Cdrain} - I_{ch\_D} - I_{Cov\_D} \tag{7-9}$$

with I_{ON} being the transistor on-current, $I_{ch\,S\,\&\,D}$ are the contributions for building up the n-channel from source and drain side, the current associated with the source or drain to gate overlay capacitance are $I_{Cov\ S\ \&\ D}$, $I_{Csource}$ is the source junction displacement current and finally I_{Cdrain} being the desired displacement current associated with the drain to well depletion region. Building $I_D + I_S$ delivers:

$$I_D + I_S = I_{Cdrain} + I_{Csource} - I_{ch} - I_{Cov} \tag{7-10}$$

with $I_{ch} = I_{ch\_S} + I_{ch\_D}$ and $I_{Cov} = I_{Cov\_S} + I_{Cov\_D}$. The contribution of the source junction capacitance can be neglected for the reference device, since the body remains on ground potential and the depletion region is not modulated. Following the above current definitions, the gate current I_G is now equal to:

$$I_G = I_{ch} + I_{Cov} \tag{7-11}$$

and adding I_G to $I_D + I_S$ yields the desired displacement current associated with C_{drain}. For the reference device, all necessary currents are plotted in the right half of Figure 7-14, whereby $I_D + I_S$ is plotted as its absolute value.

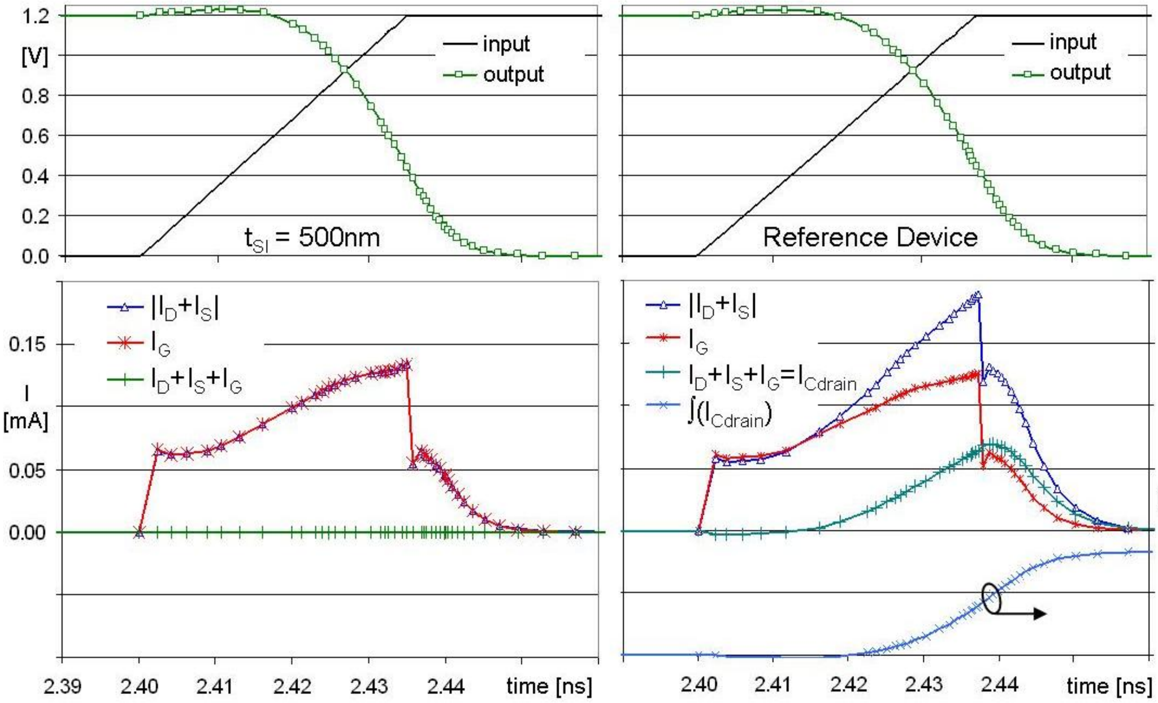

Figure 7-14: Simulated drain and source and gate terminal currents for the reference and 500 nm inverter, compared to the input and output signal

The displacement current I_{Cdrain} was integrated for the reference device and the current sums up to ΔQ ≈ 1.24 fAs for the full discharge from 1.2 V to zero regarding the drain voltage. Dividing ΔQ by ΔV delivers a rough estimate of C_{drain} ≈ 1.03 fF. Capacitance voltage simulations of the same structure deliver C_{drain}(1.2 V) ≈ 1.16 fF (see Figure 7-16), supporting the correctness of the extraction method.

The same procedure shows only negligible difference between $|I_D + I_S|$ and I_G for the 500 nm thin inverter. Due to the high well-resistance (R_{well} > 1 MΩ), the displacement current dissipating through R_{well} is reduced to a level only being of importance throughout the much longer idle (or relax) period of ≥ 1.2 ns but not effecting the immediate switching process.

Therefore, the necessary positive body charge to enable discharging of C_{drain} by $\Delta V_{Cdrain} = V_{dd} - \Delta\varphi_{b\,A} = 0.68$ V has to be delivered by C_{source}, as shown by the simulation results in Figure 7-15.

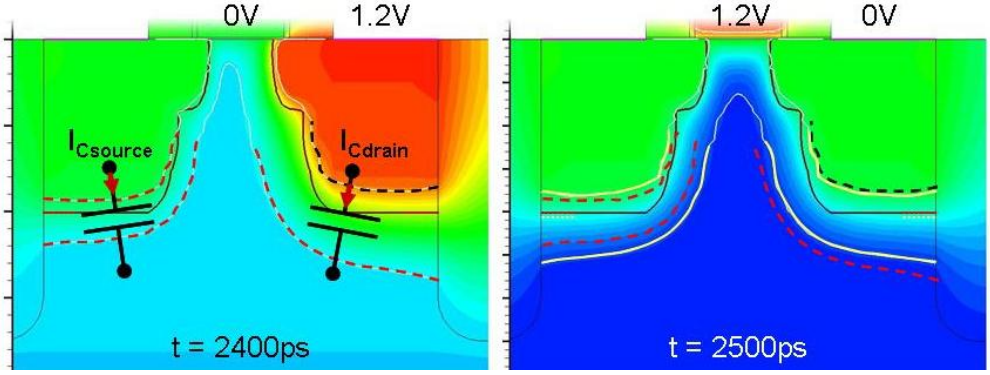

Figure 7-15: Simulated electrostatic potential of a 500 nm thin n-FET, being part of an Inverter undergoing the transition from input low to high; the dotted lines mark the end of the depletion regions

The left half shows the n-FET immediately before the transition begins, whereas the right part shows the same FET 100 ps later when the transition is ended. The surroundings of the depletion regions on the left have been superimposed to the right image (dashed lines). The modulation of the depletion regions are symmetrical, only with opposite directions from drain to source and also the resulting displacement currents are almost identical, adding to zero with $I_D + I_S$ (Figure 7-14). Consequently, the dynamic body potential is a result of the relaxation process and the capacitive voltage divider of C_{drain} and C_{source}. Both capacitances have the same size and doping profiles, hence showing the same voltage dependence which was simulated for the reference FET and is depicted in Figure 7-16.

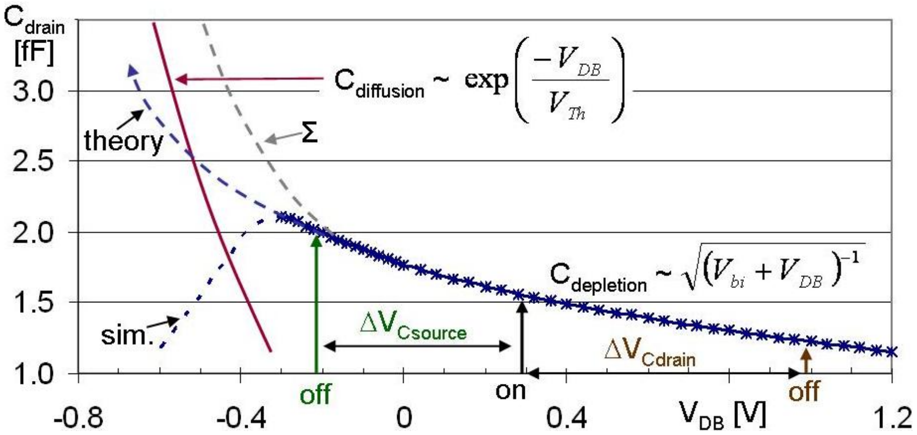

Figure 7-16: Simulated capacitance voltage curve of the drain to well junction (blue) and the theoretical (only qualitative) contribution of the diffusion capacitance of the forward biased diode (red)

The drain junction is always biased in reverse direction (depletion, with $V_{DB} > 0$), hence having relatively small values from 1.16 fF up to a maximum of 1.77 fF with $V_{DB} = 1.2$ V and 0 V respectively. The simulation failed to extract the correct capacitance for $V_{DB} < -0.35$ V due to the increasing forward current caused by the turn-on of the diode. The depletion capacitance would theoretically increase to infinity with V_{dd} approaching the built-in voltage (note that $V_{bi} < 0$ with the chosen polarity). The diffusion capacitance shows exponential dependence and becomes predominant with V_{dd}.< -0.5 V.

The situation in the thinned n-FET is also shown in Figure 7-16. The arrows mark the voltage difference across the junctions for the n-FET on- and off-state. The values of the body potential are equivalent to the simulation results shown in Figure 7-12. In on-state, both junctions are in parallel and are biased by $\varphi_b \approx -0.28$ V. In off-state, the source diode is driven in forward direction by $\varphi_b \approx 0.22$ V whereby the voltage drop across C_{drain} is reduced to 1.2 V - 0.22 V ≈ 0.98 V. The calculation of the resulting capacitance (C_{drain} in series to C_{source}) delivers ≈ 0.78 fF and ≈ 0.76 fF for the n-FET on- and off-state respectively. Comparing these values to the capacitances, as present in the reference device (with V_{DB} = 1.2 V and 0 V), shows a resulting capacity reduction of -56% in on-state and -34% in off-state. The average of 45% is very close to the 43% reduction of ΔV_{Cdrain} discussed earlier and underlines the theory that a substantial part of the speed gain can be associated with the drain capacitance reduction.

In the next step, t_{Si} was reduced to 450 nm whereby R_{well} was increased by one order of magnitude to 0.1 GΩ in accordance to the well-contact resistance measurements off the single FETs (resulting in an effective well-resistance of 18 MΩ for the n-FET and 16 MΩ for the wider p-FET). All previously discussed model alignments according to the initial body potential and normalized rise and fall times were repeated and yielded the following model input: t_{rise} = 34.8 ps, t_{fall} = 22 ps, $\varphi_{b\ n\ start}$ = 255 mV and $\varphi_{b\ p\ start\ =}$ 1.425 V. Based on these settings, the simulated speed gain was measured to be slightly above 9%, which marks the maximum positive effect achieved with the utilized simulation model.

Figure 7-17 shows the simulated n-FET body potential of the 450 nm inverter (blue) in comparison to the 500 nm results (green) and one curve simulated with reduced V_{dd} (0.8 V, orange). As already discussed in the previous part, with R_{well} entering the MΩ regime, the well-current is reduced to a level where it is too small to affect the body potential during the transition. Consequently, the major body voltage modulation $\Delta\varphi_{b\ A}$ shows only very weak dependence on R_{well} and remains almost constant going from 500 nm to 450 nm (both with V_{dd} = 1.2 V) whereby R_{well} was increased by a factor of 10. But R_{well} strongly affects the relaxation effect in-between two transitions, marked with $\Delta\varphi_{b\ B}$ during the n-FET off-state and $\Delta\varphi_{b\ C}$ during on-state.

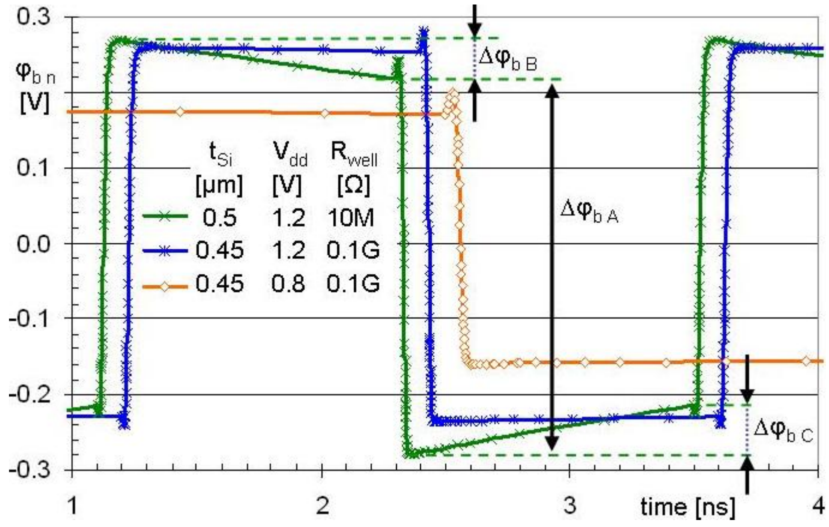

Figure 7-17: Simulated n-FET body potential for different values of t_{Si}, V_{dd} and R_{well}

The resulting body potential at the beginning of a switching event ($\varphi_{b\ n\ max}$ in Figure 7-12), and therefore the positive effect causing the V_t reduction, increases with R_{well} due to the reduced relaxation. $\varphi_{b\ n\ max}$ increases from 0.244 V to 0.282 V by ≈ 38 mV with R_{well} being increased from 10 to 100 MΩ. But the positive effect saturates with this R_{well} value since $\Delta\varphi_{b\ B\ \&\ C}$ are already reduced to < 10 mV and even with $R_{well} > 10$ GΩ (which is clearly above the extracted well-contact resistance measured at the single FETs) $\varphi_{b\ n\ max}$ does not increase any further and the speed gain does not exceed 10%.

The orange curve was simulated at $V_{dd} = 0.8$ V, with $\tau_{osc} = 6.14$ ns according to the RO data in Table 7-1, and clearly shows the supply voltage dependence of $\Delta\varphi_{b\ A}$. Reducing V_{dd} by 33% from 1.2 V to 0.8 V causes a reduction of $\Delta\varphi_{b\ A}$ from ≈ 0.5 V to ≈ 0.33 V, equivalent to -34%, indicating a linear dependence. Later discussed electron beam probing measurements on UtS devices (10.3) will support this trend, also showing linear dependence between the detected EBP signal and V_{dd} during dynamic operation.

In summary, the simulated FIB thinned devices show an increased speed because of the reduced V_t at the beginning of the transition and due to the reduced displacement current associated with C_{drain}, resulting in a reduced output load. But the simulated results remain clearly below the measured performance boost of the real RO (A), e.g. measured 18% compared to simulated 9% at $t_{Si} = 450$ nm and $V_{dd} = 1.2$ V. The discrepancy between reality and simulation will be discussed in the following.

7.3.4. Marginalities of the simulation model

A part of the difference between the real and the simulated results may be due to misalignment of the underlying FET models according to their dynamic performance and body bias sensitivity. Furthermore, some effects which have been identified based on the single FET experiment could not be implemented into the simulation model yet, but will be discussed quantitatively later on.

Due to the many unknowns with respect to the implemented doping profiles and interface conditions, the FET models were only aligned to fit the initial performance of the later FIB thinned single FETs. But since the performance alteration of the inverters in dynamic operation is strongly influenced by the modulated body potential, the body sensitivity of the simulation model was also checked and the results are shown in Figure 7-18, in comparison to the reference FET data.

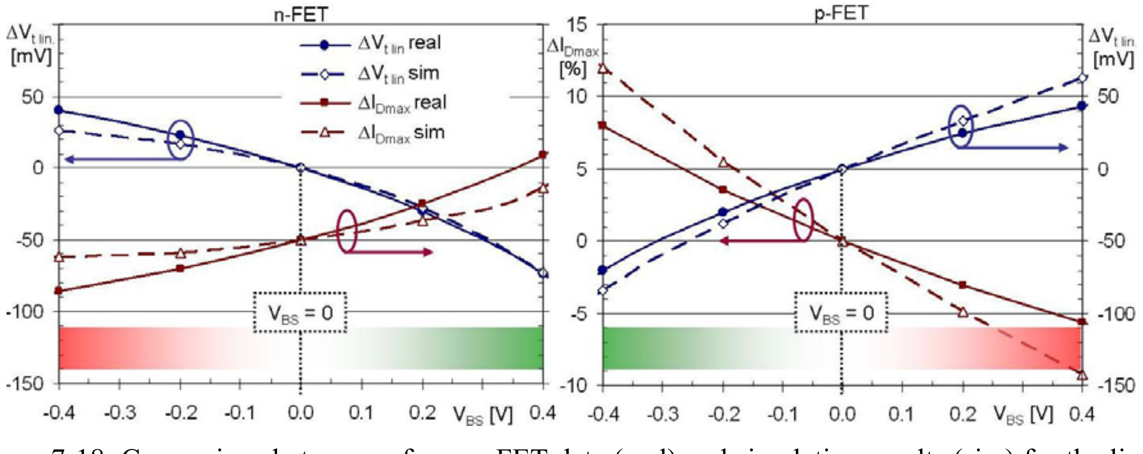

Figure 7-18: Comparison between reference FET data (real) and simulation results (sim) for the linear threshold voltage and maximum current (I_D with $V_{DS} = V_{GS} = \pm 1.2$ V) as a function of V_{BS}

All simulation results show the same trends as the real FETs, but the numbers partly differ substantially. The n-FET simulation model underestimates the body sensitivity resulting in a too low V_t decreases, hence a too low current increase. But since the p-FET tends to overestimate the V_{BS} related effects, the body sensitivity mismatches almost equalize each other and can not be held responsible for the insufficient speed gain.

A second problematic point originates from the initial alignment of the reference inverter performance, explained in (7.3.1). Only the extrinsic load capacitance $C_{L\,sim}$ was used to match the measured average inverter delay with $C_{L\,sim}$ = 13.5 fF, which is approximately 22% above the theoretical oxide capacitance of the two gates in parallel (C_{ox} = 11 fF). But the evaluation of the inverter input capacitance (based on the simulation of the 3 µm thick reference inverter, results plotted in Figure 7-19) reveals a critical discrepancy.

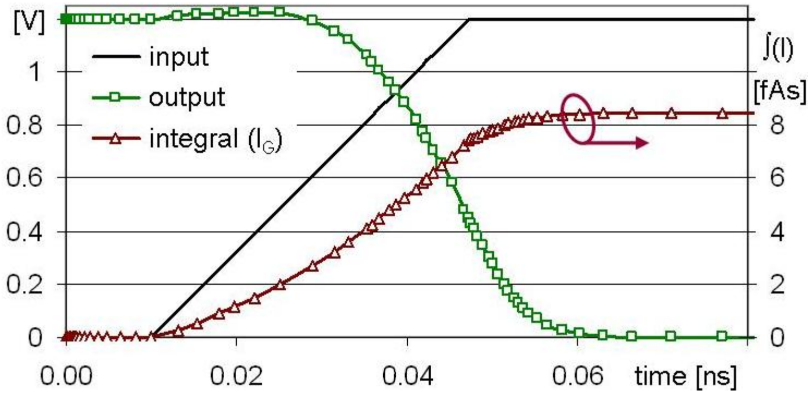

Figure 7-19: Integral over the gate current of the simulated reference inverter compared to the input and output signal

The integral over the full gate current ($I_{G\,n} + I_{G\,p}$) measures the charge difference to be $\Delta Q_G \approx$ 8.45 fAs for the full transition from V_{GS} = 0 V to 1.2 V. Dividing ΔQ_G by ΔV_{GS} delivers the equivalent constant gate capacitance $C_G \approx$ 7 fF, being almost 50% less than the implemented load capacitance $C_{L\,sim}$ which was expected to reflect the capacitive load of the next following stage.

This strong mismatch can be linked to the DC FET model alignment once again: All gate oxide and interface imperfections had to be excluded because the relevant physical characteristics were unknown and also beyond the scope of this work. Based on the given doping profiles and geometries, the poly doping had to be reduced to rather low values to shift the threshold voltage of the devices into the desired range. The finally implemented gate doping levels of $N_A = N_D = 6 \cdot 10^{18}$ cm^3 cause poly depletion effect in the applied gate voltage range, explaining the very low extracted C_G value, whereby earlier CV measurements on dedicated structures of the same 120 nm technology node did not show poly depletion. Consequently, the gate capacitance is simulated much too low.

In addition, the resistance of the poly gate was not taken into consideration within the current model. Based on the gate geometry and a typical sheet resistance of silicided poly structures of approximately R'$_p$= 10 Ω□, the full resistance of one p-FET gate (from one end to the other) can be calculated to:

$$R_{G\_p} = R'_p \cdot \frac{W_G}{L_G} = 10\Omega \cdot \frac{3.1\mu m}{0.105\mu m} \approx 300\Omega \qquad (7\text{-}12)$$

Considering that not the full gate current has to pass through the whole length of the gate, and that the n- and p-FET gates are in parallel, an input resistance of 100 Ω was chosen to evaluate the influence of the poly resistance in combination with an increased gate capacitance, realized by adding an extrinsic capacitance ($C_{G\,ext}$) to the simulated input and output capacitance.

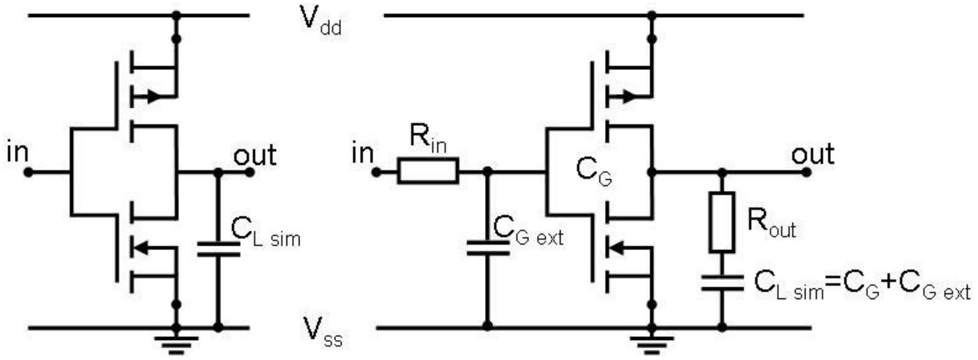

Figure 7-20: Schematics of the initial inverter model (left) and the modified model for a better alignment of in- and output capacitance and resistance (right)

The modified setup is depicted in Figure 7-20, with $R_{in} = R_{out} = 100\ \Omega$ and $C_G \approx 7$ fF, as extracted based on the gate displacement current. Now, $C_{G\,ext}$ was gradually increased to 5 fF to match the average inverter delay to the ring oscillator data. The load capacitance of $C_{L\,sim} = C_G + C_{G\,ext} \approx 12$ fF is ≈ 1.5 fF smaller compared to the initial setup.

The resulting speed gain increased by 0.5% to 9.5% with these modifications, which can be associated to the reduced load capacitance, increasing the relative effect of the simulated drain capacitance reduction, but the overall impact of the additional passive elements remains small.

7.3.5. Expected Impact of Diode Degradation

All effects discussed so far are based on the increased well-contact resistance in association with the reduction of the active material volume. The ion beam related surface degradation could not be implemented into the simulation model since the resulting material degradation is not known precisely enough, as already discussed at the end of (6.3). But the observed diode degradation measured on the single FETs and discussed in (6.2) gives solid ground for the hypothesis that defect cascades penetrate the material and start to substantially increase the diode reverse current when the remaining distance between the rear surface and the depletion region is less than 200 nm.

For the single FETs it was concluded that the resulting leakage current of the drain junction increases the static body potential by up to 200 mV (see Figure 6-22). The same effect also influences the body potential in dynamic operation. Figure 7-21 shows the n-FET of the inverter in dynamic operation. With input low and resulting output on V_{dd}, the n-FET is in off-state and $V_{DS} = V_{dd} = 1.2$ V. In addition to the simulated modulation of the body potential, associated with the loading and unloading of the drain and source capacitances, the degradation induced junction leakage (represented by $R_{leak\ R\ \&\ F}$) acts like a highly resistive voltage divider, elevating φ_b during n-FET off-state, or at least reducing the relaxation in this period.

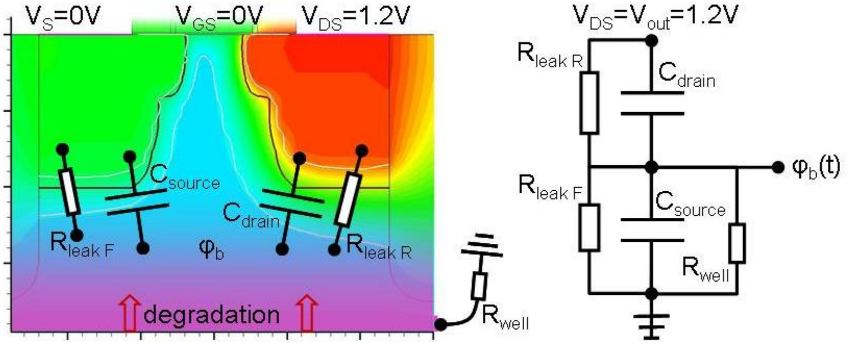

Figure 7-21: FIB thinned n-FET in off-state (left) and schematics of the diodes

Furthermore, the two leaky junctions ($R_{leak\ R\ \&\ F}$) and R_{well} are connected in parallel during the n-FET on-state (with V_{out} = V_{DS} = 0) and increase the relaxation during this period. This increased relaxation causes a less negative (higher) body potential before the transition from output low to high, and assuming $\Delta\varphi_{b\ A}$ to be constant this results in a further increased φ_b after the transition. The process is illustrated in Figure 7-22, in comparison to the simulated body potential of the 500 nm inverter.

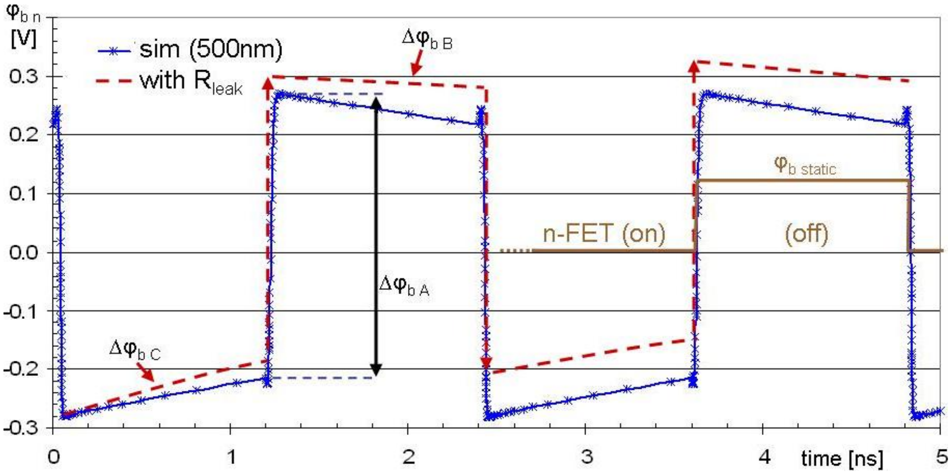

Figure 7-22: Simulated body potential of 500 nm thin inverter and the illustrated effect of the junction degradation on φ_b

The resulting increased average body potential translates into additional performance increase of the n-FET. The p-FET experiences the exact same effect, with different signs but also with increasing performance. Consequently, including the diode degradation into the simulation model would increase the simulated speed gain, bringing the simulation results closer to the measured 18% performance elevation of the real RO structures with V_{dd} = 1.2 V.

Another influence of the ion beam related material degradation is depicted in Figure 7-23. The crystalline defects in combination with the Ga contamination do not only degrade the diodes, they also substantially reduce the carrier mobility, hence increase the resistance. Even though the lateral distance between the source and drain areas is only in the 100 nm range, the lateral resistance adds additional delay to the displacement current the higher the distance to the center of the device gets. Consequently, the loading and unloading of the already reduced drain junction capacitance will be spread over time. This does not reduce the capacitance itself, but the modulation of the drain depletion region can happen slower, hence the output is loaded less during the transition.

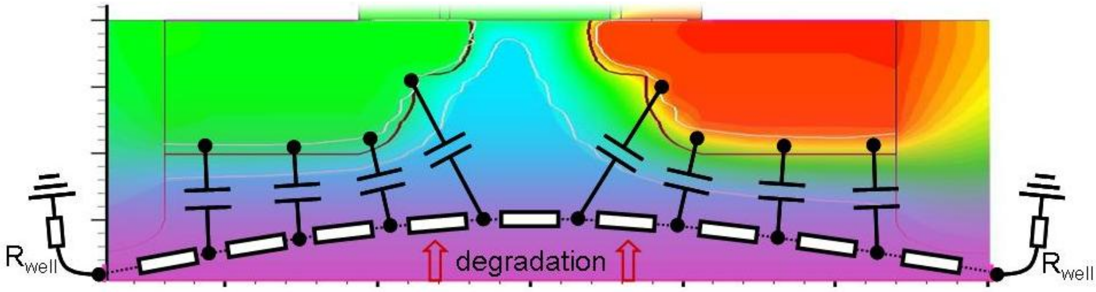

Figure 7-23: Rear surface degradation causing increased lateral resistance within the FET body

7.3.6. Weight of the Different Effects & Supply Voltage Dependence

The theoretical discussion of the most important intrinsic inverter parameters, with respect to the resulting dynamic performance, delivered a linear dependence between the average stage delay τ and the so called switching resistance R_{sw} multiplied by the sum of all capacitances attached to the output node (Formula 4-21). In accordance to the modified inverter model, as shown in Figure 7-20, the theoretical equation for τ can be simplified to:

$$\tau \approx R_{sw} \cdot \left(C_{drain\_n} + C_{drain\_p} + C_{L\_sim} \right) \qquad (7\text{-}13)$$

For simplicity, the simulated voltage depended drain junction capacitances (as shown in Figure 7-16 for the n-FET) will be approximated by a constant average capacity per unit area, measured at 0.4 V reverse bias to $C'_{drain\ n} \approx 3$ fF/µm^2 and $C'_{drain\ p} \approx 3.66$ fF/µm^2. With the dimensions of RO (A) (L_d = 180 nm, W_n= 2.7 µm and W_p= 3.1 µm), the average junction capacitances are $C_{drain\ n} \approx 1.5$ fF and $C_{drain\ p} \approx 2$ fF. In combination with $C_{L\ sim}$ = 12 fF, the full load adds up to 15.5 fF.

The simulation results in 7.3.3 have shown a 45% reduction of $C_{drain\ n}$ as a result of the thinning process. Assuming a similar reduction of $C_{drain\ p}$ would reduce the output load to ≈ 13.9 fF which is equivalent to -10%. Following equation (7-13), this directly translates into 10% delay reduction.

Even though all capacitances are voltage dependent, similar evaluations for V_{dd} = 0.8 V and 1.6 V did show only marginal voltage dependence of the capacitance reduction with 45±3%. As a result, the capacitive effect contributes with approximately 10% to the speed gain for all thee supply voltages.

These 10% are already sufficient to explain the measurement results with V_{dd} = 1.6 V. The stronger performance increase measured with nominal and reduced supply voltage must therefore be linked to the floating body potential in combination with the diode degradation. As discussed in 7.3.5 (only for V_{dd} = 1.2 V), the increased diode leakage current shifts the average body potential $\overline{\varphi_b}$ from the simulated values close to zero into the positive. Figure 7-24 shows the simulated n-FET body potential for the three different supply voltages. With V_{dd} = 1.6 V, the high modulation amplitude of $\varphi_{b\ n}$ (> 0.65 V) would result in $\varphi_{b\ n\ max}$ > 0.3 V in case of a symmetric signal (with $\overline{\varphi_b} \approx 0$). But the turn-on of the source well diode limits $\varphi_{b\ n}$ to approximately 0.3 V. As a result, the whole signal is shifted towards the negative and the average body potential also becomes negative. Also the degraded diodes in the real RO are not believed to shift $\overline{\varphi_b}$ into the positive, explaining the missing positive effect of the still present V_t reduction at the start of the n-FET dominated transition.

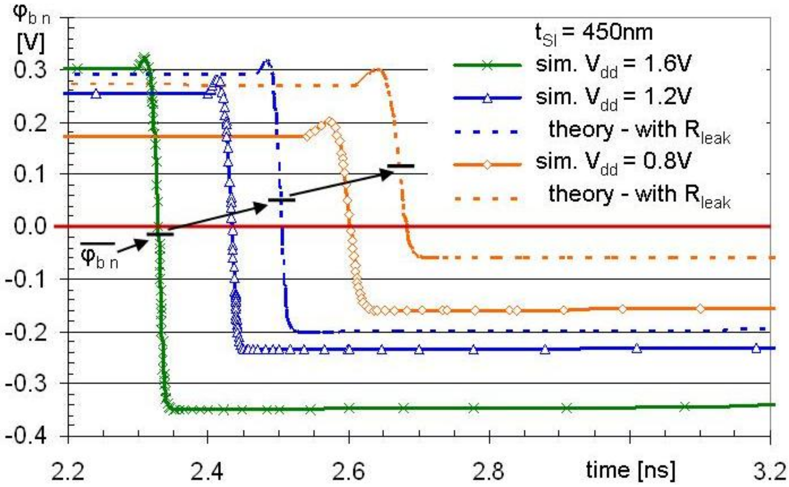

Figure 7-24: Simulated n-FET body potential for V_{dd} = 0.8 V, 1.2 V and 1.6 V and shifted curves resulting from the diode degradation theory

With V_{dd} = 0.8 V, the amplitude of the body potential modulation is linearly reduced to ≈ 0.33 V and the simulated maximum value remains clearly below the onset of the source well diode. Consequently, the drain diode leakage results in a stronger increase of the average body potential and $\varphi_{b\ max}$ is assumed to be close to the values simulated with V_{dd} = 1.2 V and 1.6 V, as illustrated by the dashed lines in Figure 7-24. As a result, V_t remains on a low level, similar to higher supply voltages, whereby the same V_t reduction has a much stronger performance impact due to the disadvantageous V_t / V_{dd} ratio, also being responsible for the low start performance (f(0.8 V) < 40% f(1.2 V)). Furthermore, the negative impact of the reduced body potential during transistor on-state is strongly reduced for V_{dd} = 0.8 V since the average body potential results to be clearly above zero.

Further proof for this theory and additional evidence for the presence of one more leakage based speed enhancement effect will be given by additional experimental results, discussed in the following. Especially the strong decline of the performance increase measured when the remaining silicon thickness approaches STI height has to be discussed based on the additional experimental data at the end of this chapter.

7.4. NAND based RO (B)

The second ring oscillator RO (B) is comprised of 65 NAND gates, as illustrated in Figure 7-25. Two inputs of the 3-way NAND are constantly connected to V_{dd}, reducing the logical function to a simple inverter.

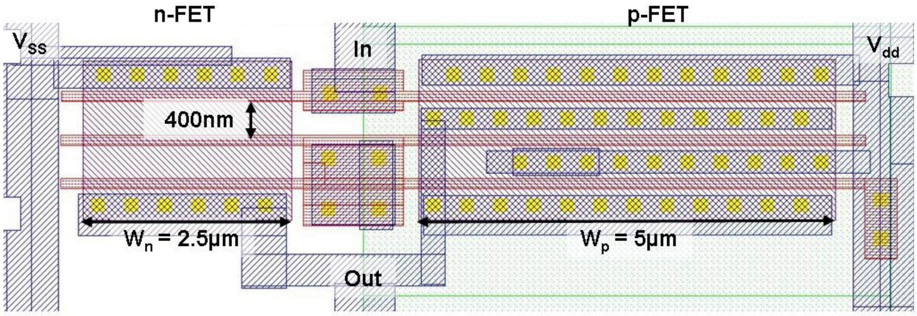

Figure 7-25: Layout of one NAND cell of RO (B), with two inputs being constantly connected to V_{dd}

A schematic cross section through the n- and p-FET part of this NAND is depicted in Figure 7-26. For the p-FET side, the two additional transistors are constantly switched off and only the one which is directly connected to the input undergoes the switching, following the input signal. Therefore, only the drain to well junction of the most right p-FET contributes to the output load capacitance. In contrary to that, the two additional n-FETs are constantly switched on and are connected in series to the first n-FET. Since they are both directly wired to the output node, their full drain/source regions contribute to the load capacitance and furthermore, the two channel regions result in an additional resistance, limiting the current of the switching n-FET. Figure 7-26 shows the simplified schematic of the NAND gate including the on-resistances of the constantly on n-FETs ($R_{\text{on n}}$) and the single capacitive contributions of the relevant diffusion areas (C_d).

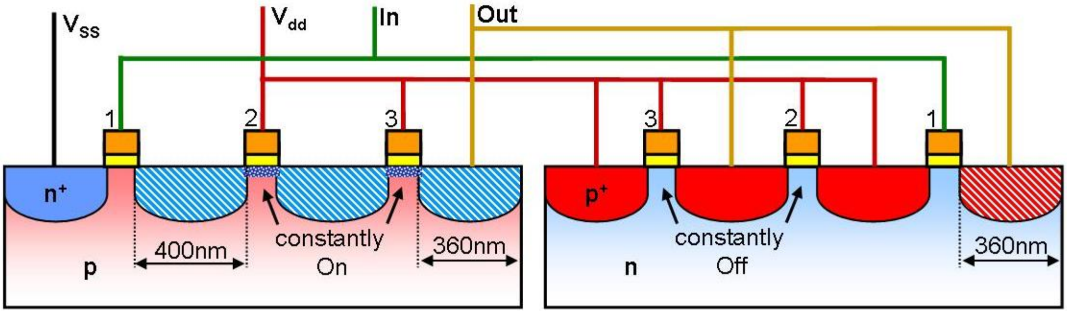

Figure 7-26: Schematic cross section through a NAND cell of RO (B). All drain/source regions, contributing to the capacitive load at the output node are filled with striped pattern.

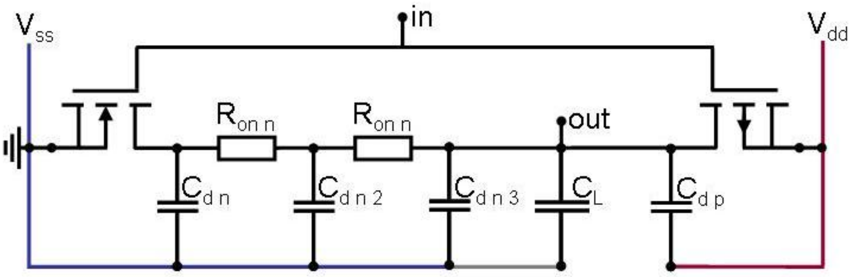

Figure 7-27: Schematic of a single NAND gate, including the most important parasitic capacitances and the on-resistance $R_{\text{on n}}$ of the two constantly on n-FETs

The normalized on-resistance of the n-FETs can be approximated by use of the maximum current of the single n-FET (Figure 6-10).

$$R'_{on\_n} \approx \cdot \frac{V_{DS}}{I_{D\max}} \cdot W = \frac{1.2V}{5.7mA} \cdot 10\mu m \approx 2.1k\Omega \cdot \mu m \tag{7-14}$$

With $W_n = 2.5\ \mu m$ $R_{\text{on n}}$ results to be ≈ 840 Ω. Similarly, the capacitances can be derived based on the simulated capacitance per unit area to:

$$C_{dn} = C_{dn2} \approx {}^{3\,fF}\!\big/_{\mu m^2} \cdot 0.4\mu m \cdot 2.5\mu m = 3\,fF \tag{7-15}$$

$$C_{dn3} \approx {}^{3\,fF}\!\big/_{\mu m^2} \cdot 0.36\mu m \cdot 2.5\mu m = 2.7\,fF \tag{7-16}$$

$$C_{dp} \approx 3.66\, fF \Big/ _{\mu m^2} \cdot 0.36 \mu m \cdot 5 \mu m \approx 6.6\, fF \qquad (7\text{-}17)$$

Together with $C_{ox} \approx$ 14.2 fF, the full capacitive output load is increased to approximately 29.5 fF. As a result of the higher load and the increased series resistance ($2{*}R_{on\,n}$), the initial ring frequency at the nominal supply voltage is only about 40% compared to RO (A). Figure 7-28 shows the ring output frequency as a function of the silicon thickness. Similar to RO (A), the ring performance remains stable until $t_{Si} \approx$ 700 nm. But in contrary to the first results the performance increase is much stronger with a maxima of approximately 25%, 30% and 40% for V_{dd} = 1.6 V, 1.2 V and 0.8 V respectively. Additionally, the speed enhancement remains at around 20% at STI height, where it had already dropped to below 10% at RO (A).

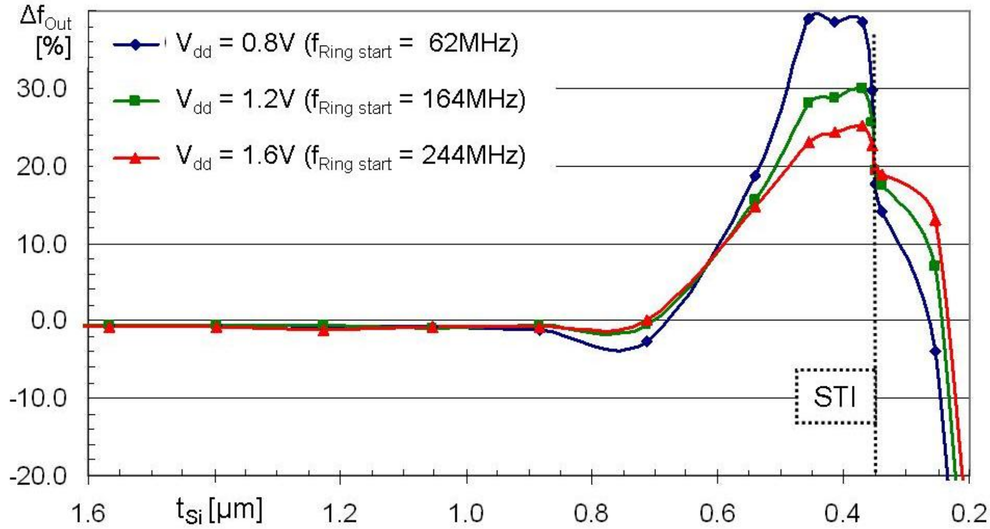

Figure 7-28: Progression of RO (B) frequency, measured at three different supply voltages and drawn as a function of t_{Si}

Most of the response to the FIB thinning can be linked to the reduction of the high initial capacitance. Assuming the drain capacitances to decrease similarly as simulated for RO (A), by -45%, this already results in a reduction of the full load capacitance by 23.4%. The real capacitance reduction can be expected to be even stronger (≈ 25%) due to the strong asymmetry between the contributing drain area (2*0.4 µm + 0.36 µm) and the source area (only 0.36 µm), leading to a lower resulting capacitance when connected in series, compared to the symmetric case at RO(A). Also, the body potential modulation can be expected to be stronger because of the same reason.

Similar to RO (A), almost the full +25% of performance measured with V_{dd} = 1.6 V can be linked to the reduction of the drain capacitance, whereby the additional +15% with V_{dd} = 0.8 V must be driven by the body potential related V_t reduction.

But in contrary to this test structure (RO B), where only input 1 carried the input signal, all three NAND inputs would most likely be subject to signal changes in a real device. The already discussed transition initiated by input 1 is the slowest possible transition of this gate due to the high sum of the contributing drain capacitances which also offers the highest relative acceleration (≈ 25%) due to the strong relative capacitance reduction.

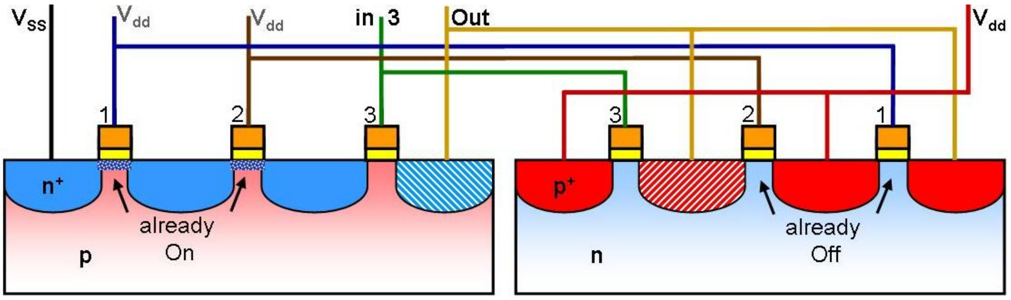

Figure 7-29: Modified situation for RO (B) when input 3 drives the transition

Considering input 3 to drive the switching, the contribution of the drain capacitance to the initial delay would already be strongly reduced as illustrated in Figure 7-29. Consequently, an output transition initiated by input 3 is substantially faster (for the unmodified/reference device) which also reduces the achievable FIB induced speed gain.

The positive conclusion is that the initially slowest transition of such multi-input gates, which is most likely to be the origin of a timing issue, shows the strongest positive response to the FIB procedure. On the other hand, this adds more complexity to the prediction of the FIB induced speed gain when working on more complex circuitry.

7.5. NOR based RO

The last RO (C) is designed with 65 3-way NOR gates (Figure 7-30). Again, two inputs are constantly wired to a fixed potential, V_{SS} in this case. The cross sectional view through such a NOR gate is depicted in Figure 7-31. Only one n-FET drain area contributes to the output load whereby three p-FET diffusions are now part of C_L (as highlighted with the striped pattern filling and depicted in the schematics in Figure 7-32).

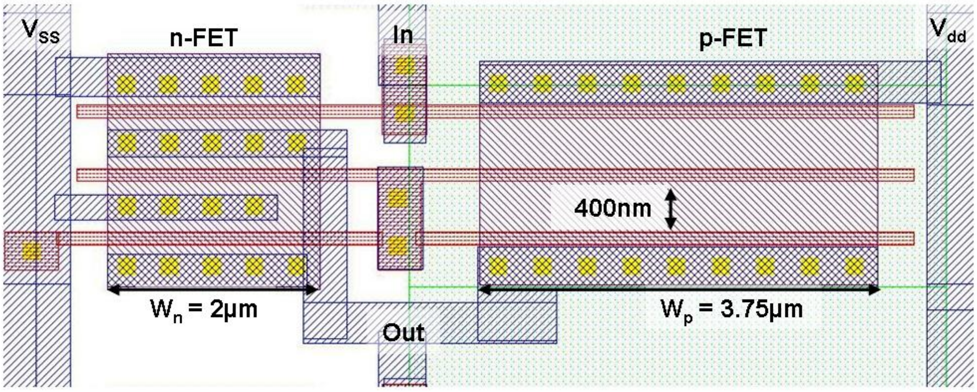

Figure 7-30: Layout of one NOR cell of RO (C), with two inputs being constantly connected to V_{SS}

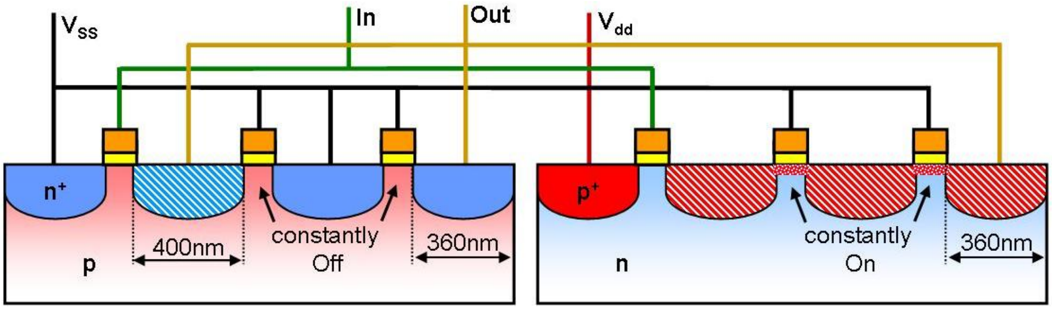

Figure 7-31: Schematic cross section through a NOR cell of RO (C). All drain/source regions, contributing to the capacitive load at the output node are filled with the striped pattern.

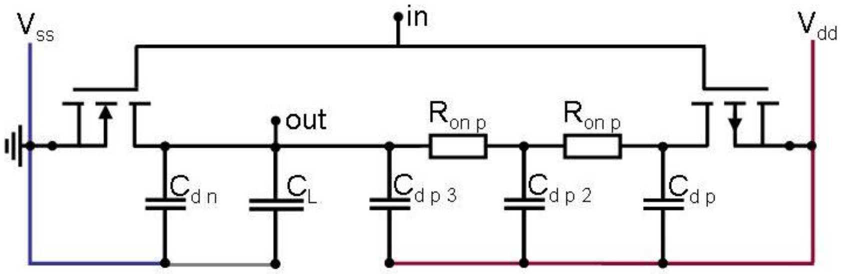

Figure 7-32 Schematic of a single NOR gate, including the most important parasitic capacitances and the on-resistance $R_{on\ p}$ of the two constantly on p-FETs

Similar to RO (B), the values of the equivalent passive components were calculated to $R_{on\ p} \approx 1$ kΩ, $C_{d\ p} = C_{d\ p\ 2} \approx 5.5$ fF, $C_{d\ p\ 3} \approx 4.9$ fF, $C_{d\ n} \approx 2.4$ fF and together with $C_{ox} \approx 10.9$ fF, the full output load sums up to ≈ 29.2 fF. This capacitive load is almost the same as with RO (B), but due to the reduced transistor width compared to RO (B) the initial ring frequency is further reduced to below 30% of RO (A).

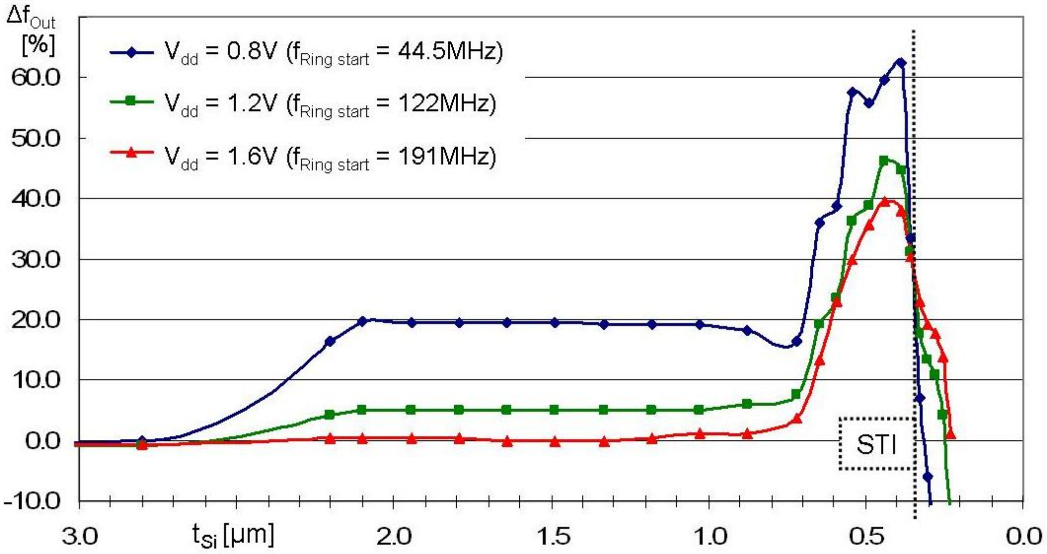

Figure 7-33: Progression of RO (C) frequency, measured at three different supply voltages and drawn as a function of t_{Si}

In contrary to the first two test structures, RO (C) already shows a strong and very V_{dd} dependent performance increase with $t_{Si} \leq 2.2$ µm. The reason for the different behavior lies in the unusual design of the ring where the n-well, covering/surrounding all p-FETs, has no direct connection to V_{dd} as would normally be present in CMOS technology. The resulting effect is illustrated in Figure 7-34. All p-FET sources are constantly wired to V_{dd}, hence under start conditions (no FIB thinning) all source diodes are biased in forward direction and increase the well potential to almost V_{dd}. The reverse current through the drain junctions and through the vertical n-well to p-well junction can be neglected. And, since all individual p-FET bodys are low ohmically connected together and share the static body potential of ≈ V_{dd}, the ring shows the expected initial performance.

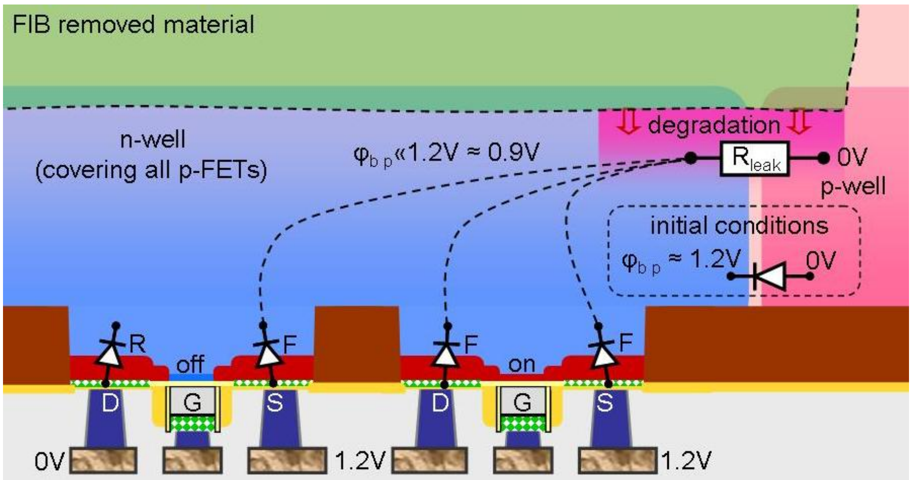

Figure 7-34: Effect of well to well leakage caused by the FIB related material degradation

As soon as the FIB thinning reaches the lower end of the n-well, the vertical pn junction between the two wells becomes strongly degraded. The result is illustrated by the additional resistor (R_{leak}), representing the FIB induced leakage current. One n-well surrounds all p-FETs with a perimeter of ≈ 400 µm. The ring transistors, including their drain and source junctions, are not degraded until t_{Si} falls below 1 µm. Consequently, the leakage of the vertical pn junction in combination with the intact drain source diodes ($V_{bi} > 0.8$ V) causes the static n-well (or n-body) potential to drop substantially in this thickness range (2.2 µm < t_{Si} < 1 µm) and the measured performance increase can be associated to the resulting V_t reduction. This effect will later be referred to as static V_t reduction because the underlying leakage current of the vertical junction is static, hence not being affected by the switching or operation state of the single transistors.

The FIB induced leakage of the drain junctions measured on the single FETs have shown only relatively weak voltage dependence, which allows to assume that the resulting difference between source and body potential in the modified RO (C) will be close to the same value for all three supply voltages ($\Delta\varphi_{b\,p} \approx$ 1.2 V - 0.9 V = 0.3 V, as given in Figure 7-34, note this is not a calculated value but rather an assumption). Consequently, the measured performance increase of ≤ 1%, ≈ 6% and ≈ 20% with V_{dd} = 1.6 V, 1.2 V and 0.8 V respectively gives further proof for the theory that a certain V_t reduction has a much stronger effect the lower V_{dd} becomes.

The FET junctions start to degrade at $t_{Si} < 0.8$ µm. As measured on the single FETs (Figure 6-6) the degradation does not only increase the reverse current but also the forward current, hence reducing the voltage drop across the junction for a fixed forward current value. Considering a constant well to well leakage, this reduces the resulting body bias which can be seen at the slight reduction of the positive effect at $t_{Si} = 0.7$ µm.

Reducing t_{Si} further will separate the individual FET bodies from each other, suppress the influence of the static leakage of the vertical pn junction and initiate all earlier discussed effects similar to RO (A) & (B). The main contributor of the resulting final speed gain will again be dominated by the reduction of the drain junction capacitance. But since no additional conclusions can be drawn, these effects will not be discussed anymore.

7.6. Well to Well Leakage and Summary of all Identified Effects

Going back to the more conventionally designed and therefore more representative RO (A), the results of RO (C) can be utilized to understand the overly strong decrease of the measured speed gain when t_{Si} approaches STI height. Since both wells are directly wired to V_{ss} and V_{dd}, the FIB induced well to well leakage can dissipate via the low ohmic well-contacts as long as $t_{Si} > 700$ nm and the average body potential remains unaffected. But as measured for the single FETs, the well-contact drastically increases for the n- and p-FET in the final thickness range (700 nm > t_{Si} > 350 nm). With the utilized RO (A), this leads to the situation depicted in Figure 7-35 where the average body potential of the n- and p-FET becomes influenced by the series connection of $R_{well\ p}$, R_{leak} and $R_{well\ n}$.

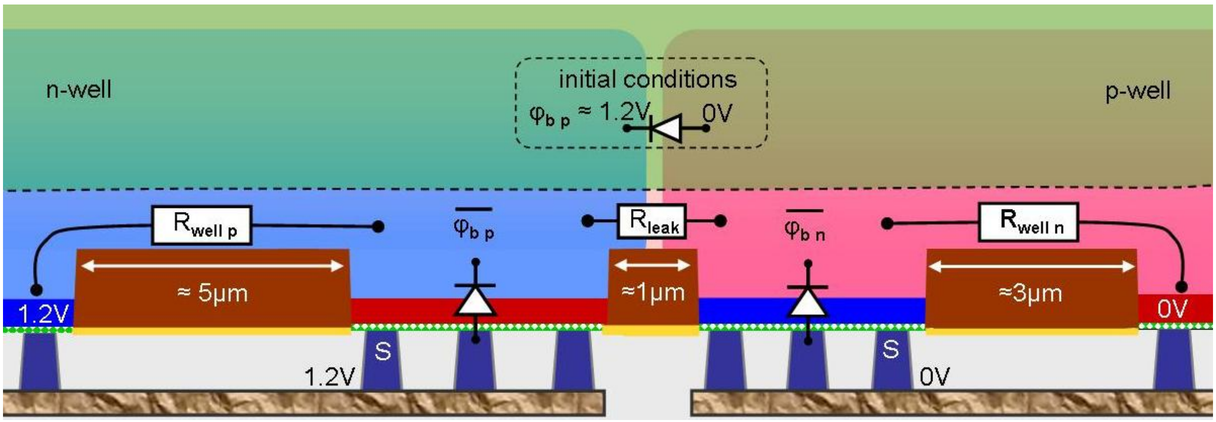

Figure 7-35: Effect of well to well bias, illustrated based on the geometries of RO A

Now, the well to well leakage current causes a voltage drop across the two well resistors pushing the average body potential into the positive for the n-FET, and below V_{dd} for the p-FET. This additional average body voltage shift has the same positive effect through static (not switching related) V_t reduction as already discussed for the drain source diode degradation in 7.3.5. But in contrary to the drain source diode degradation, the well to well leakage effect is suppressed when t_{Si} reaches STI height because all three resistors increase into the MΩ and become irrelevant for the resulting floating body potential. The utilized test structures did not allow measuring the well to well leakage and the effect can not be isolated from the expected average body potential increase associated with the drain source leakage, but a second closer look into the results measured at RO (A) allows a rough quantitative separation of the identified effects.

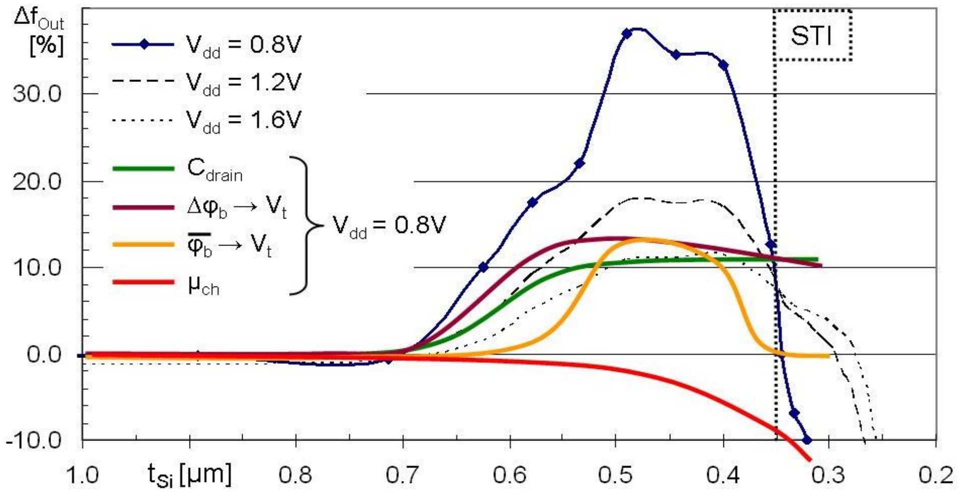

Figure 7-36: Frequency change in %, measured on RO A and expected contribution of the individual effects for V_{dd} = 0.8 V

With the speed optimized (folded) design of RO A, the drain junction capacitance marks only ≈ 23% of the full capacitive load. Consequently, the reduction of C_{drain} only contributes to approximately 10-12% to the speed gain, with the full effect still being present at STI height. With V_{dd} = 0.8 V, also the dynamic nature of φ_b has a strong effect ($\Delta\varphi_b$ in Figure 7-36). Since φ_b follows the drain potential of the individual transistor, V_t is reduced for the n- and p-FET at the beginning of their dominated transitions (input rising edge for the n-FET and falling edge for the p-FET). The dynamic V_t reduction is believed to contribute with 12-14% ($\Delta\varphi_b \rightarrow V_t$). Both effects ($C_{drian}$ and $\Delta\varphi_b$) are already initiated at relative low R_{well} values (>10 kΩ) because of the comparably high currents during the inverter transition. The average floating body potential is increased later (t_{Si} < 600 nm) since the low well to well and drain diode leakage current does not noticeably affect the body potential before R_{well} increases into the MΩ. But, in between 500 to 400 nm, the additional static body potential increase (and associated V_t reduction, $\overline{\varphi_b} \rightarrow V_t$) can be expected to contribute by clearly more than 10% (≈ 13% in the illustration) based on the experimental results of RO (C), with 20% speed gain only related to the well to well leakage at higher t_{Si} values. Approaching STI height, the well to well leakage effect becomes fully suppressed and in combination with the increased degradation of the FET channel (as quantitatively discussed in 6.3, see Figure 6-19 with -8% at STI height), represented by the channel mobility (μ_{ch}), the speed gain is reduced to approximately 10% with t_{Si} = 350 nm.

In summary, the dynamic performance increase results from a complex balance between increased junction leakage, increased well-resistance and related capacitance and V_t reduction. The effect is strongly supply voltage dependent due to the increased sensitivity of the dynamic performance to V_t changes with reduced V_{dd}. The positive effect of the capacitance reduction is almost constant (not V_{dd} dependent), whereby the maximum contribution is directly related to the ratio between the full load being attached to the output node and the initial sum of all later reduced (drain) junction capacitances. Furthermore, all these effects show a new level of design dependence - the lateral distance between n-well-contact, p-FET body, n-FET body and p-well-contact (as depicted in Figure 7-35) has a direct impact onto the quantity of the individual performance increasing effects, whereby these distances are considered being uncritical in common CMOS designs as long as design rules are not violated.

8. Applications for FIB Induced Dynamic Performance Increase

The monitored speed gain initiated by the proposed FIB thinning procedure opens a whole new application field for backside FIB in the debug of aggressively designed semiconductor products. In current and future chip designs, timing related soft fails become the predominant limitation, hence post-fabrication timing adjustments mark a critical need for future debug.

This chapter briefly reviews the established techniques for timing trimming before discussing the applicability, costs and benefits of the new FIB technique.

8.1. Established Techniques for the Debug of Timing Conflicts

The only non physical way to adjust internal chip timing after the DuT has been fabricated is by changing the capacitive load being attached to a given circuit node. In general there are two well established techniques. In both cases, the designer adds an additional performance margin to the driver of the critical node. Furthermore, additional capacitances are attached to these nodes, either as constant capacitances in combination with fuses or in the form of varactors. With the first technique, multiple fuses allow disconnecting multiple capacitances adjusting the inner loads. The latter technique makes use of varactors and is illustrated in Figure 8-1. Initially all trimming voltages V_0 to V_2 are set to the same value, hence all varactors have the same capacitance. Usually the initial voltage is chosen in a way to start with the lowest possible capacitance for C_0 to C_2. In this example, signal S_0 is too slow to guarantee correct logical operation of the following node. Once this timing conflict is located, the individual signals can be aligned by changing the trimming voltages. Here, the capacitances C_1 and C_2 are increased, hence slowing down S_1 and S_2.

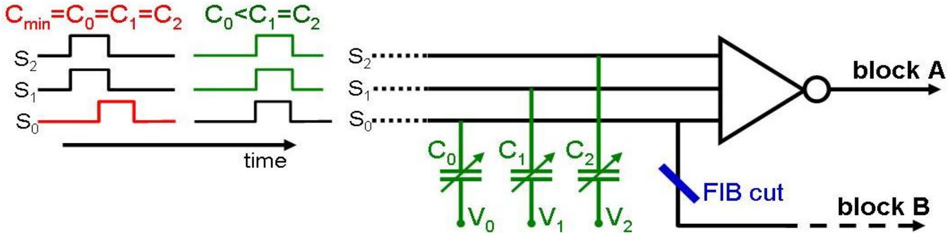

Figure 8-1: Illustration of established techniques to trim circuitry internal timing conditions

The main drawback of these techniques is the additional chip size necessary to integrate the capacitances and fuses or the varactors in combination with their control logic for programming the trimming voltages.

At unexpected fail locations, hence without such pre-designed elements, only conventional circuit edits can help reduce the load by disconnecting unnecessary branches. In the depicted example, the delay of S_0 could be reduced by cutting the interconnect towards block B. Doing so may allow to test block A at the desired speed but the DuT does not have its full functionality anymore. Consequently, this method would not allow the fabrication of fully functional engineering samples which are often highly desired for customer prequalification or early prototyping.

8.2. Timing Adjustment by FIB Thinning Procedure

With the demonstrated FIB induced dynamic performance increase it is now possible to trim the chip internal timing without the use of pre-designed elements or destructive/load reducing CE. The upper part of Figure 8-2 shows another example of a timing conflict, in this case caused by a typical circuit edit, as is often necessary during design debug. The additional 20 ps delay introduced by the (high ohmic) FIB connection propagate unchanged through the following gates and result in a timing violation on the next clock-relevant logical unit. Since there is no additional load to be disconnected, the classical CE approach would fail on this problem.

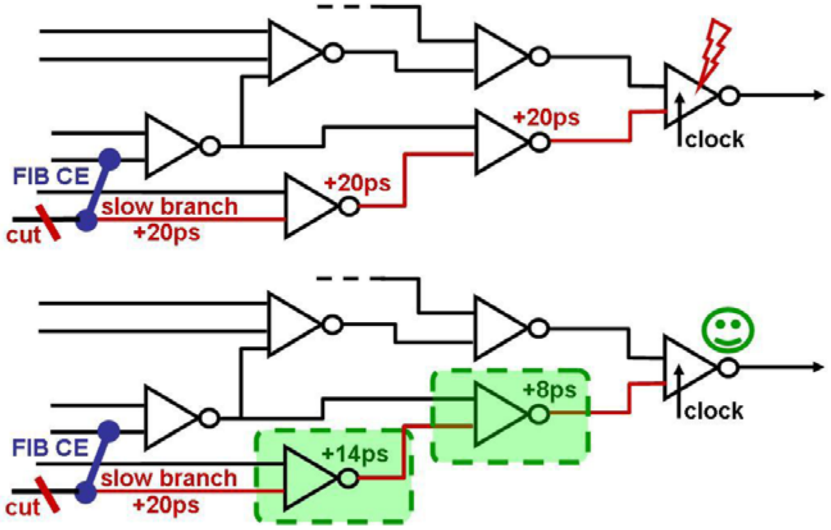

Figure 8-2: Circuitry showing a timing conflict caused by a FIB CE without (upper) and with the FIB induced speed enhancement locally applied to two gates, restoring correct timing conditions (lower)

Applying the FIB thinning procedure to selected gates of the slow branch allows increasing their speed individually. In the depicted example (lower part of Figure 8-2) a delay reduction of 6 ps per stage is assumed to be achieved by the FIB procedure. Increasing the speed of two gates would already be sufficient to regain correct timing conditions.

The major benefit of this technique lies in its ubiquitous applicability, allowing speeding up any desired gate on a given DuT. Furthermore, due to the unsurpassed co-planarity between FIB trench bottom and device level (discussed in 3.2.2), the FIB induced speed enhancement can be applied to full circuit blocks covering several 100 µm^2 at the same time.

One disadvantage of the proposed technique originates from the complexity of the FET internal effects being responsible for the measured performance increase. As discussed in the previous chapter, the resulting speed gain of a single gate is a strong function of the supply voltage, the individual circuit design (FO, inverter, NAND, NOR, 2-way, 3-way…) and the remaining Si thickness at the end of the thinning procedure. For gates with multiple inputs also the order in which the individual inputs are changed will have an effect on the quantity of the FIB induced speed gain. Consequently, the finally achieved positive effect on a more complex modified circuitry may not be easily predicted.

Given the ability to measure the timing modification in-situ (directly in the FIB vacuum chamber) would allow to subsequently modify several gates of a critically slow branch until the desired timing is established. But with today’s chip complexity, number of IOs and necessary clock speed to drive a DuT at its maximum speed, the in-situ measurement would require establishing a dedicated, rather complex and cost intensive setup.

In contrary to prior results on 180 nm inverter chains [29] & [30], the utilized 120 nm ROs did not show their speed gain maxima at the technology given thickness of the STI shapes (350 nm) but in the thickness range from 400 to 500 nm. Aiming at the maximum effect would therefore enforce to control the remaining Si thickness during the FIB thinning with less than ±50 nm precision which is not possible with the current setup. R. Jain et al. [31] proposed a technique using the coaxial optics of the OptiFIB to determine the remaining Si thickness base on different fringe patterns captured at two slightly different light wavelengths. Implementing two laser light sources (with the necessary finite wavelengths) into the optical path of the system may yield the desired precision but was not yet tested.

Using the STI bottom as the endpoint of the FIB thinning yields less performance gain but is much simpler to achieve with the commercially available FIB systems. Furthermore, at STI height the speed gain is reduced to a level which is almost equivalent to the effect of the drain capacitance reduction. In contrary to the complex origin of the maximum delay reduction present at higher thickness values, the quantity of the capacitance related performance increase is easily predictable, marking another advantage of using the STI endpoint.

One very critical point is the increased self-heating which can become a severe problem when chip areas with high power density are exposed to the extensive FIB thinning. Theoretically it is possible to seal the FIB thinned areas with a thin FIB deposited isolation layer and fill the remaining voids with any heat conducting material afterwards. This may allow reestablishing the necessary thermal stability for high speed testing.

Additionally, reducing V_{dd} can be advantageous in many ways: At first, the power consumption drops with $\approx V_{dd}^2$, drastically reducing the self-heating problem. Secondly, also the maximum test speed is reduced with similar dependence as the power consumption, which would allow for a less complex test setup (especially beneficial considering the proposed in-situ timing measurements). Finally, also the maximum delay reduction introduced by the FIB thinning is strongly V_{dd} dependent, showing approximately twice the effect when V_{dd} is reduced from nominal (1.2 V) to 0.8 V on the utilized 120 nm technology RO test structures. But, the DuT may not show the same timing issues with reduced V_{dd} compared to the nominal supply voltage, limiting the usefulness of the V_{dd} reduction.

In summary, the proposed FIB thinning procedure is the only technique allowing trimming the chip internal timing in the absence of dedicated pre-designed elements and without load reducing CE. Consequently, this technique can deliver fully functional engineering samples, which are often desired for pre-qualification of a product or early prototyping, and could not be achieved yet. The complex multi-parameter dependence of the measured speed gain requires precise characterization in combination with circuit simulation (not necessarily physical) for any new technology node being worked on. Only this can allow precisely predicting the resulting timing modifications in a real DuT, necessary to plan such modifications.

9. Backside Circuit Modifications based on UtS

The most important available backside CE contact methodologies are depicted below (Figure 9-1) and will be discussed in the following.

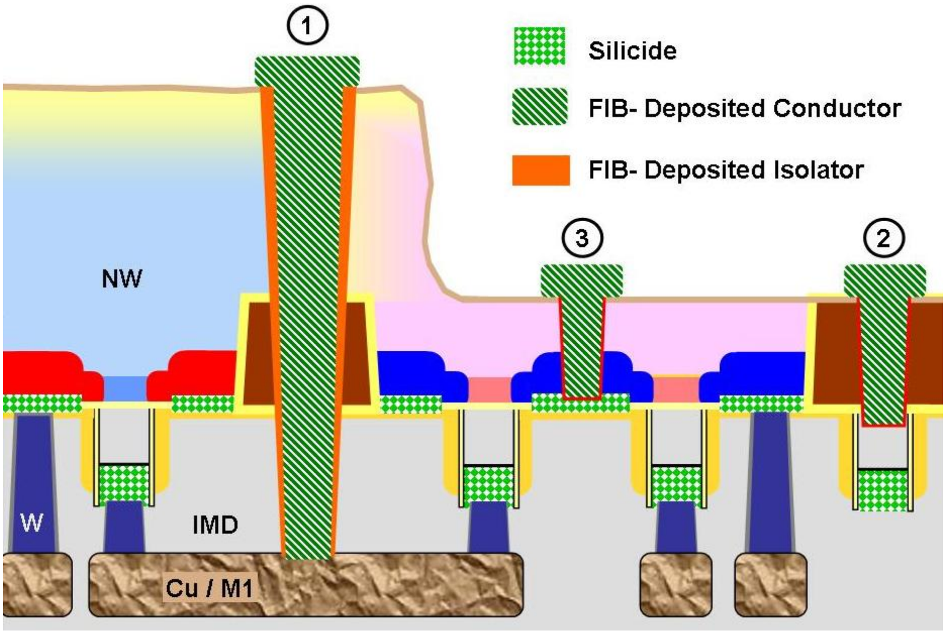

Figure 9-1: Illustration of three different methodologies to establish an ohmic contact to an inner circuit node

(1) Contact to Metal Interconnects

(2) Contact to Poly (CtP)

(3) Contact to Silicide (CtS)

Except for the contact to metal, which is the conventional way of doing backside CE, all other methodologies are only possible starting from UtS level. The Contact to Poly has not been fully characterized yet and will only be discussed very briefly. In comparison to the conventional- and the CtP approach, CtS has shown superior contact properties, offers a wide range of applications, and will be discussed in more detail.

Note that all cut operations still have to be carried out on interconnect level. Cut operations are usually less critical, since less susceptible for the unintended creation of short-circuits.

9.1. Contact to Metal Interconnects

The direct contact to interconnect lines is the most common routine for today's FIB CE. Compared to frontside CE, the backside CE approach offers much better access to the lower metal levels. The contact from interconnect metal (Aluminum or Copper) to the FIB-deposited conductor (compound of Platinum, Tungsten or Molybdenum) is purely ohmic and has a low interface resistance. But in highly dense chip areas, the access to targeted metal lines can be blocked, or critically limited by active structures (e.g. transistors), poly routing or even lower metal levels. Figure 9-2 illustrates a typical CE situation. In this case, the first input and the output of a 3-way NAND gate are only covered by electrically unused area (STI, covering all non-diffusion area, not having an extra mask layer, hence not explicitly shown in the image) being easily accessible and could be contacted or cut by uncritical (low aspect ratio) FIB operations. The other two inputs (in this example constantly wired to V_{dd}) are connected via metal 2, being fully blocked by the transistor itself and cannot be addressed using this strategy. For a real CE in dense chip area the access is always strictly limited, enforcing to mill high aspect ratio node access holes. Aspect ratios of > 25/1 have successfully been demonstrated [16], but even though the hole can be milled correctly, the high aspect ratio of the conductor fill bares severe problems, ending in critically increased via-resistance [17]. It is also likely that a targeted circuit node has to be cut and reconnected for the desired CE, which is only possible having two separate HAR node access holes or one bigger opening, further limiting this approach.

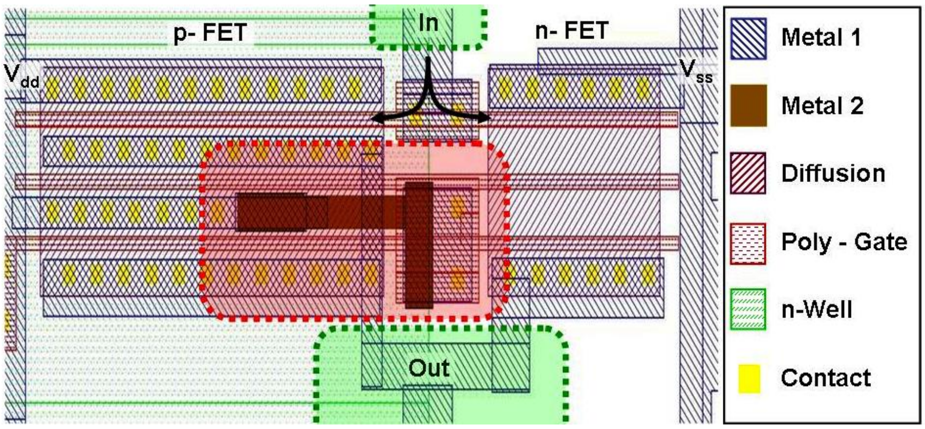

Figure 9-2: CMOS NAND layout with critical and uncritical areas in terms of CE

9.2. Contact to Poly (CtP)

One common design rule in recent technologies says that all contacts to poly are not allowed to be fabricated on top of a gate structure. Consequently, poly gates always have to have an additional area besides the transistor. As shown in Figure 9-3, these areas are big enough to securely establish contacts to poly. One challenge with any FIB milling procedure is to find the right endpoint for the operation. Starting with UTS and milling an access hole to a poly structure, the etch operation passes through the STI, followed by an oxide layer, both highly isolating and appearing dark in a FIB image. In contrast to these isolators, a grounded poly structure provides sufficient endpoint information when exposed by the ion beam. A proof of concept was published in [32], but the contact properties have not been investigated further.

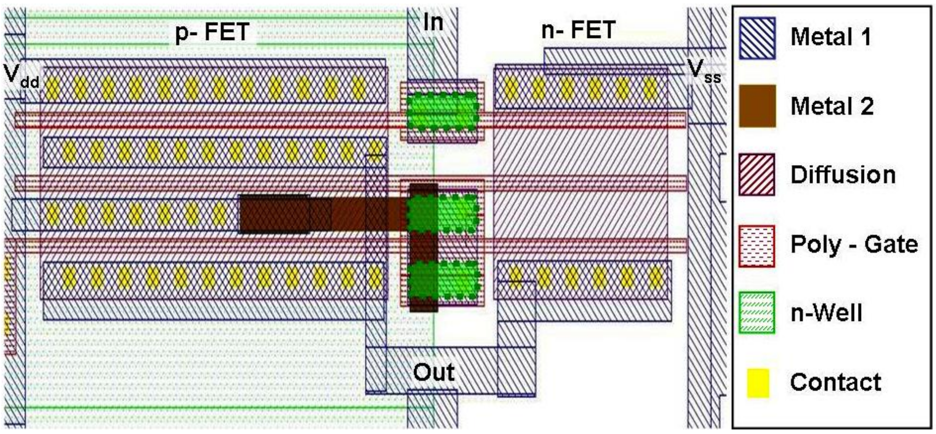

Figure 9-3: CMOS NAND layout with potential areas for Contact to Poly

9.3. Contact to Silicide (CtS)

The main benefit of CtS is that it allows addressing every circuit node, hence every signal on a given chip. The Silicide is a highly conductive 30-40 nm thick layer which covers all source/drain areas in recent technology generations. The dashed green lines in the layout example (Figure 9-4) mark suitable contact locations, where the contact would be established to the transistor source/drain areas and the underlying metal at the same time. Furthermore, CtS can make contact to diffusion areas which have not been contacted in the original fabrication process. This cannot be achieved with any other frontside FIB process. The bright green rectangles cover such locations in between two poly-gates. This allows CE to be done directly to the device level for the first time.

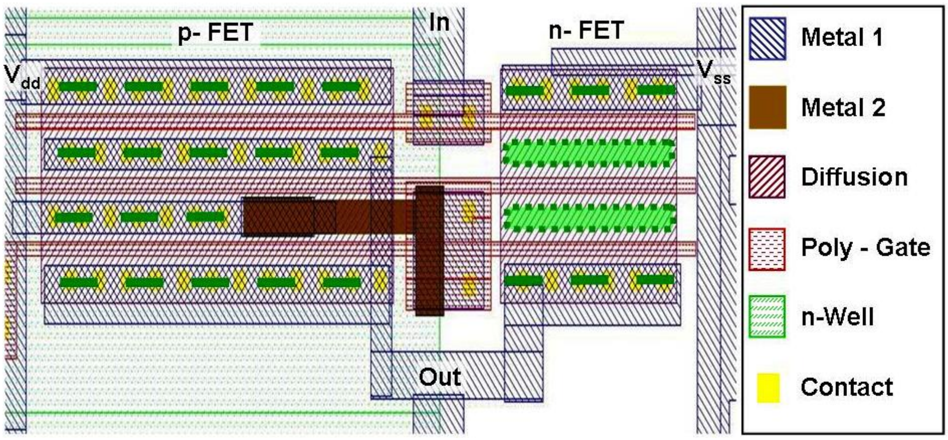

Figure 9-4: CMOS NAND layout with additional areas, only accessible by CtS

The advantages for CE are illustrated in Figure 9-5. The left side shows a typical CE, including cutting and reconnecting a circuit node. Following the standard backside CE approach the shown solution would be the only possible way due to the high integration density. But the close proximity between the neighboring M1 line and the deposited via-fill would pose a high risk to the success of the CE.

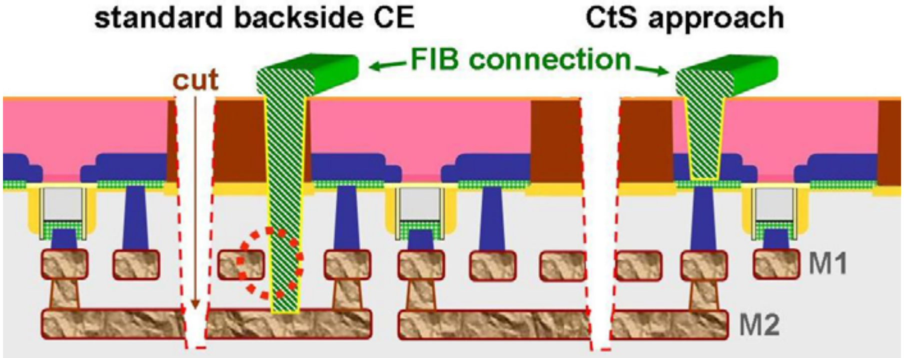

Figure 9-5: Comparison between conventional and CtS based CE

For the slightly modified situation depicted in the right half of the image, the same CE would be impossible in the conventional way. But with CtS only the cut operation has to be done on the metal line and the CE could be performed without problems.

Trying to stop FIB milling operations in a 30-40 nm thin layer would usually be regarded as very challenging, hence having a low success rate. Starting from UtS level overcomes this problem since the start Si thickness is always equivalent to STI height (≈ 350 nm), offering reproducible milling conditions with low aspect ratio holes. Figure 9-6 shows a low-angle section through a source/drain area. The contrast between CoSi and p^+ diffusion is faint, but by use of an endpoint detection tool (a typical endpoint curve is shown in the right part of the image), endpoint control proves to be uncritical.

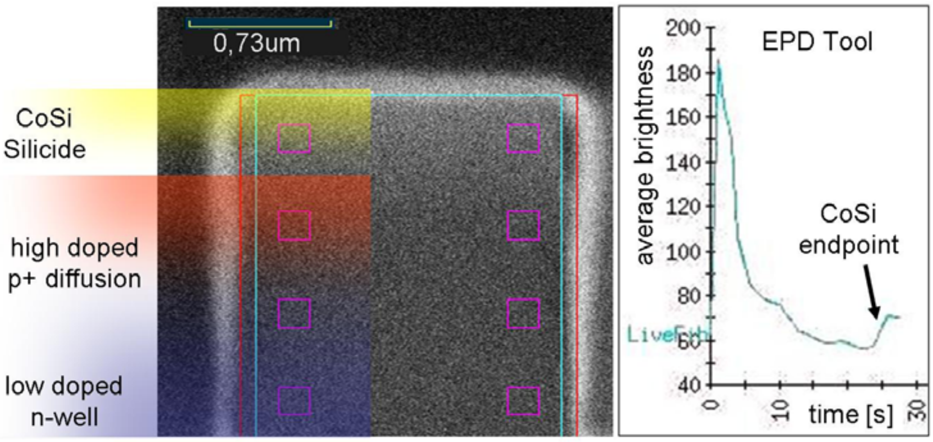

Figure 9-6: FIB image of source/drain area, gradually thinned down to the Silicide layer (left) and endpoint detection signal of atypical milling operation stopping on the CoSi Silicide level (right)

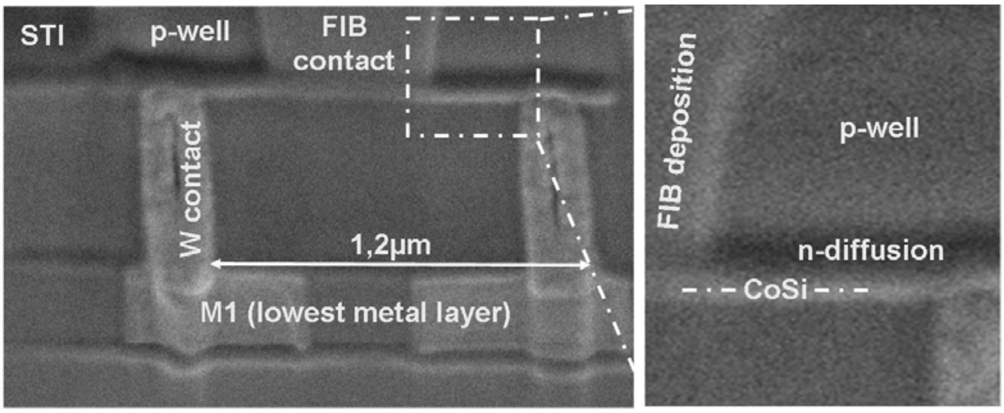

Figure 9-7: SEM image of a cross-section through a CtS structure

A cross section through a CtS is shown in Figure 9-7, where the bottom of the FIB deposited conductor ends perfectly flat on top of the Silicide layer.

The contact properties were measured by use of an atomic force prober (AFP). The results in Figure 9-8 prove the formation of a reproducible and purely ohmic contact with less than 400 Ω each (60 Ωµm^2 with the used contact size of 0.8x0.2 µm^2). The contact resistance scales linearly with the inverse of the contact area. Consequently, editing on bigger devices which drive more current allows increasing the contact size/reducing the contact resistance, offering additional CE solutions where the conventional approaches may already fail.

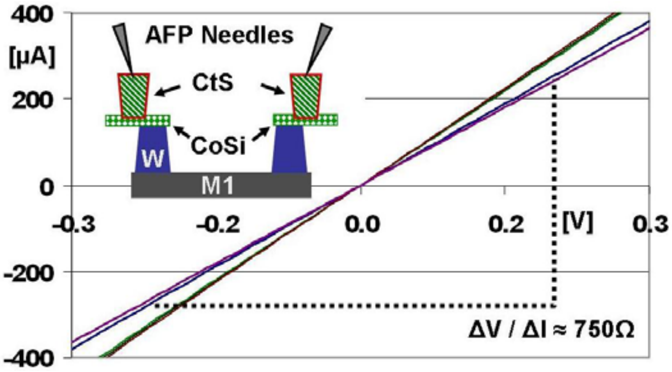

Figure 9-8: Current vs. voltage curves of various CtS, measured by backside AFP (always two CtS with 0.8x0.2µm^2 contact size connected in series via M1)

In addition to the degrading impact of the UtS formation discussed earlier, the CtS fabrication further stresses the device. The critical points are highlighted in Figure 9-9, compared to a contact to a metal interconnect.

As is well known from basic semiconductor theory, a metal contact on moderately doped material results in a Schottky-diode, hence showing rectifying behavior. Considering the contact to metal, the node access hole passes through the highly doped well-material. With a typical well-doping level of ≈$5 \cdot 10^{18}$ cm^{-3} at the peak position, the FIB deposited conductor would form such a Schottky-diode (point (3) in Figure 9-9). Due to the voltage levels present in common CMOS technology (n-well on V_{dd}, p-well on V_{ss} and any interconnect metal switching between either of those) any such Schottky-diodes (also when milling through p-well) would be either unbiased or biased in reverse direction. But the resulting reverse current may still be high due to the relatively high well-doping levels and the poor sidewall conditions (amorphous layer). This additional leakage current may pose a risk to the success of the CE and consequently the sidewalls of such access holes may have to be isolated all the way down to the STI (4), requiring additional process steps.

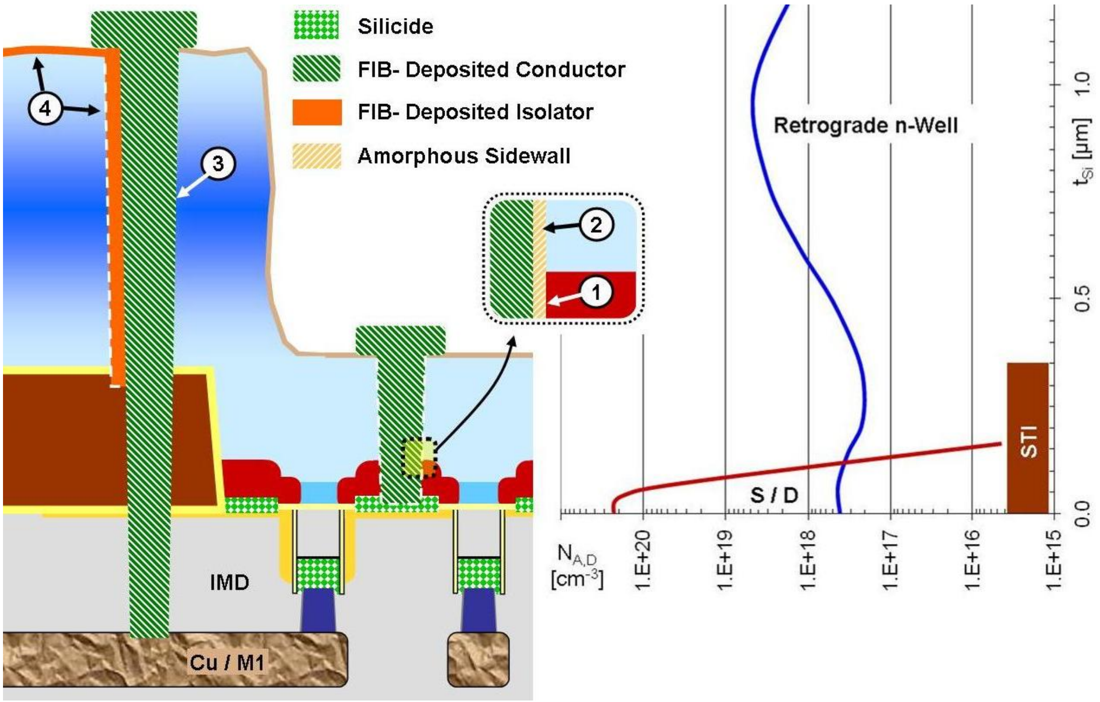

Figure 9-9: Comparison between a contact to metal starting on well level and a CtS (left) drawn to the same scale as a typical S/D and well-doping profile of recent technologies (right)

Milling the access hole for the CtS also creates a thin amorphous layer on the side walls, degrading the pn-junction of the drain to well diode. To evaluate the impact, a single 10 µm wide p-FET structure was FIB processed and the drain to well diode was monitored all the way through. Figure 9-10 shows a FIB image of the CtS location and the captured data. The UtS area was intentionally limited to 4x4 µm^2 to keep the remaining n-well material connected still covering the majority of the device area. This allows measuring the increased reverse current ($I_{DB\ rev}$) without the superimposed influence of the increasing R_{well} (see Figure 6-6 including R_{well} in comparison). The first curve (bare Die) shows the initial diode characteristic of a reference device, measured on a low noise probe station. All other measurements were done in the FIB vacuum chamber. The diode reverse current should still be fully unaffected on n-well level, hence the first strong increase must be due to the setup leakage and should not be interpreted. As already seen in chapter 7, the diode reverse current is first elevated above the setup limit within the final stage of the UtS formation. Compared to the reference diode, $I_{DB\ rev}$(-1.2V) increases from $\approx 7 \cdot 10^{-14}$ A to $\approx 3 \cdot 10^{-10}$ A. Regarding that only about 20% of the full drain area was thinned to STI level this marks an increase by a factor of $\approx 2 \cdot 10^{4}$ in the UtS area. The subsequent milling of the 1x0.5 µm^2 access hole down to the Silicide layer resulted in a further increase of $I_{DB\ rev}$(-1.2V) to $\approx 9 \cdot 10^{-8}$ A associated with the degradation of the sidewalls. Compared to the initial diode reverse current this marks an increase by a factor of $\approx 10^{7}$. This strong increase may be critical, especially for analog applications, but compared to the maximum on-current of such FETs (> 300 µA/µm) this additional leakage current (< 0.1 µA) is not expected to pose a problem for CE in a CMOS environment.

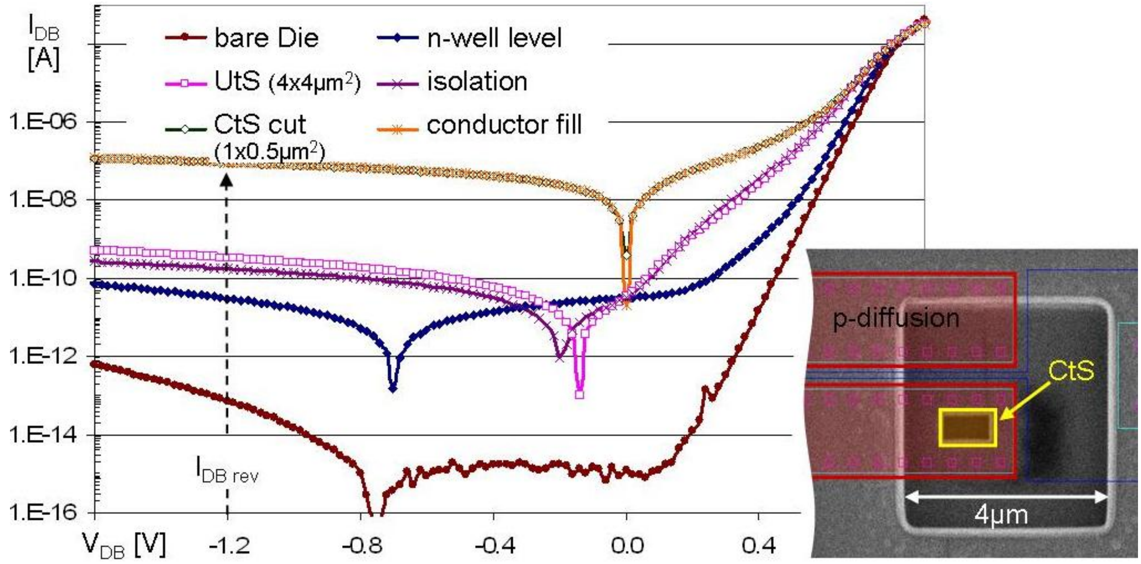

Figure 9-10: Drain to well diode characteristics, measured after different necessary steps for CtS

The subsequent filling of the access hole with FIB deposited conductor had only negligible effect onto the diode leakage. The side-contact to the highly doped drain/source area (1) (see Figure 9-9) reduces the resistance to the targeted circuit node and is therefore appreciated. The resulting Schottky-diode between the conductor and the remaining FET body (2) has better rectifying properties compared to the diode between metal-fill and well (3) because of the lower doping in the relevant thickness range. Additional sidewall isolation is not necessary.

9.4. Application Examples of CtS Methodology

In the following, two application examples of CtS are shown where the contact methodology is used to prove its applicability beyond classical CE tasks. First, the small contacts were utilized for a direct probing task. A 120 nm technology RO was thinned to n-well level (left part of Figure 9-11). The ten stage divide, which reduces the internal ring frequency by a factor of 1/1024 was chosen for this experiment. Using the n-wells for local alignment, ten circular holes were milled down to STI level surrounding the p-FET part of a simple inverter. This inverter marks the input of every individual divider stage. Subsequently, all ten drain areas of theses p-FETs were connected via CtS. The high frequency signal of the ring enters the first divider and passes MP1 at its full speed. The input of the following stage carries only half the frequency and so on. The diameter of the circular access holes (3 µm) was aligned to the shape of the used active probe needle. In this way, the needle only had to be lowered into one hole and automatically made contact to the CtS on the bottom. The side wall guidance reduces the requirements regarding the positioning precision and stability of the utilized probe station substantially.

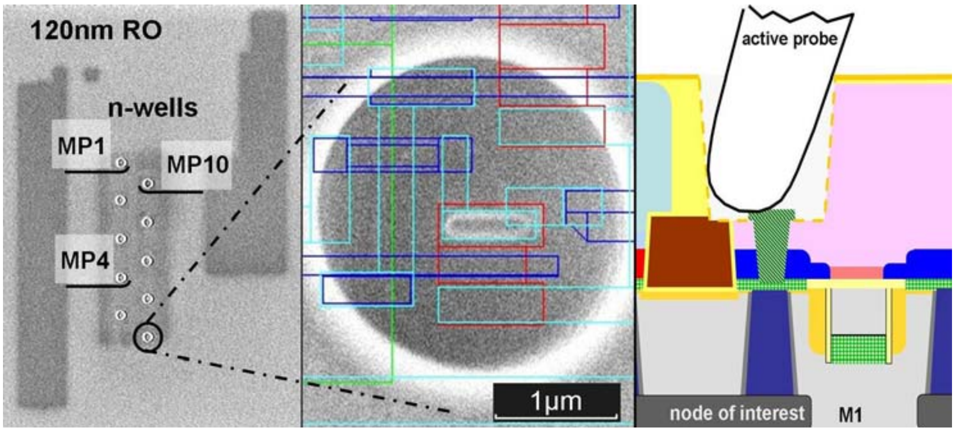

Figure 9-11: Application example for CtS: n-well image (left) of a 120 nm technology RO, with 10 circular holes; enlarged view of one hole with CtS and layout overlay (middle) and sketch of CtS as used for direct probing (right) - the trench sidewalls allow for self alignment of the active probe

Figure 9-12 shows some measurement results. The signals of MP 1, 4 and 10 are compared to the final output signal f_{out} and show the expected frequencies of $f_{X10} = 2*f_{out}$, $f_{X4} = 128*f_{out}$ and so on. Also the ring frequency itself with ≈ 0.1 GHz could be measured by use of one of the simplest active probes on the market (Picoprobe 12C-4-10, $f_{MAX} = 500$ MHz).

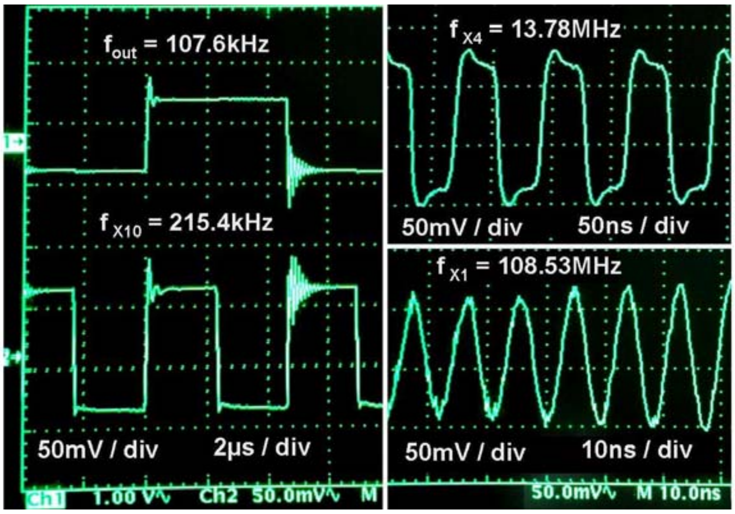

Figure 9-12: Oscilloscope snapshots with signals measured by direct probing on CtS

The second application example focuses on single device characterization. Dedicated nano-probing systems have to be used due to the very small geometries with recent technologies. These systems either use SEM to control the positioning of the probes or the probes themselves have scanning probe microscopy (SPM) capability, allowing the aligning of any individual probe based on a previously captured topology image. The latter systems are called atomic force prober (AFP). With both setups, the backend of the DuT has to be lapped down to the contact layer to gain access, automatically separating all transistors from each other. The resulting setup is depicted on the left of Figure 9-13. This technique is well established and allows the precise and reproducible characterization of the fully isolated transistors, but if the failure was located in the metal lines it is irreversibly removed with the backend. Furthermore, if the root cause can not be found amongst the initially suspected transistors, the failure analyst has no second chance to conduct additional fault localization techniques.

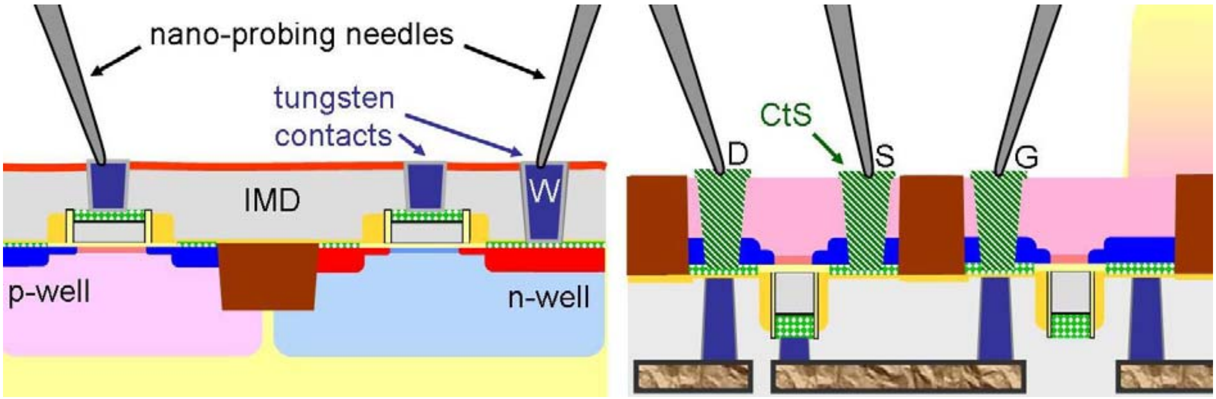

Figure 9-13: Comparison between conventional AFP (left) and backside AFP based on CtS (right)

Regarding a typical debug or FA flow, most analysis techniques have to be applied through the chip backside. Consequently, the package has to be opened on the rear and the chip is thinned to around 100 µm or less. The frontside nano-probing would now enforce the full removal of the package and the already mentioned lapping to the contact layer. In contrary to that, the backside characterization approach (right half of Figure 9-13) can be applied without further package or backend removal. The small geometries of the CtS allow contacting all necessary FET terminals. Figure 9-14 shows the FIB image of the test location and the enlarge layout of the targeted CMOS inverter. Since the backend is still untouched, the two single transistors can only be measured separately when the connection between their drain contacts is cut, as illustrated in the layout. The two drains are contacted via CtS directly on the targeted FETs, the source contacts can be established further apart since the sources are always connected to either V_{dd} or V_{ss}. The gate is contacted by CtS to the output of the former stage, which is not shown in the layout but the idea was already depicted in Figure 9-13.

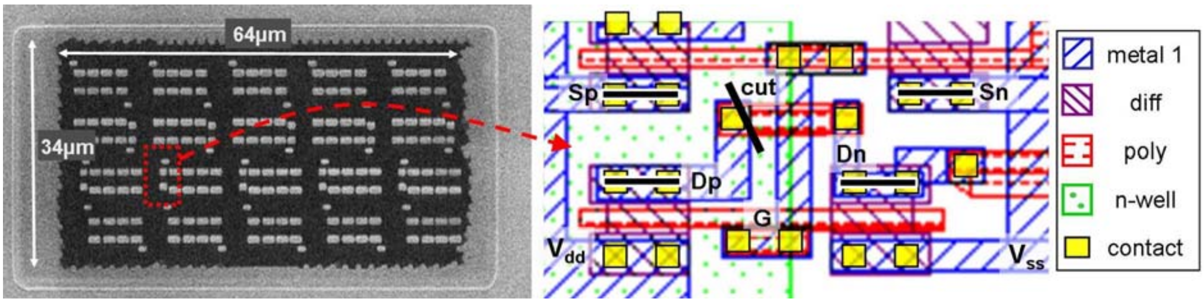

Figure 9-14: FIB image of a 120 nm CMOS logic chip readily prepared to UtS (left) and enlarge layout of edit location including the location of 4 CtS (Dn, Sn, Dp & Sp) and one cut

All contacts were fabricated with the same size of 0.8x0.2 μm^2, resulting in an ohmic contact resistance of less than 400 Ω each. This additional contact resistance is uncritical since known and reproducible. The results are shown in Figure 9-15.

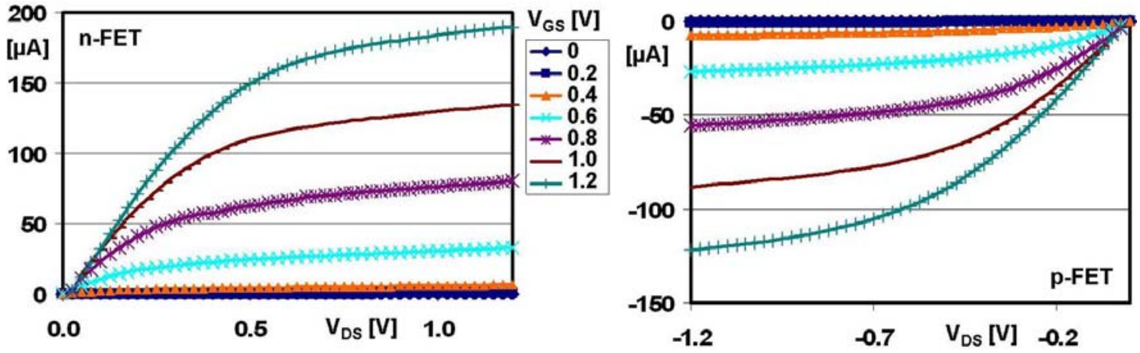

Figure 9-15: Output-characteristics of n- and p-FET (W_G = 730 nm, L_G = 160 nm), measured by AFP, contacted via CtS

This technique can not compete with the established frontside approach in terms of measurement accuracy, especially because of the degrading nature of the FIB thinning and CtS formation as discussed earlier. Nevertheless, these results prove that a rough performance evaluation is possible.

The major advantage of the backside technique is that the whole circuitry remains functional. The necessary cut to separate the targeted FETs can be bridged by only one conductor deposition in less than two minutes (connecting the two drain CtS), allowing to go back and forth with applying localization and characterization. Furthermore, once the failing FET is pinpointed a cross-sectional SEM inspection or even TEM lamella extraction can be done in the same approach directly through the chip backside with no additional preparation in between.

10. UtS as a new Platform for Circuit Analysis

With the ability of controlled and localized FIB thinning to any desired thickness, optical and electron beam based probing techniques can also benefit, according to their general applicability and maximum resolution. The general ideas are depicted in Figure 10-1 and the costs and benefits of these techniques will be discussed in the following.

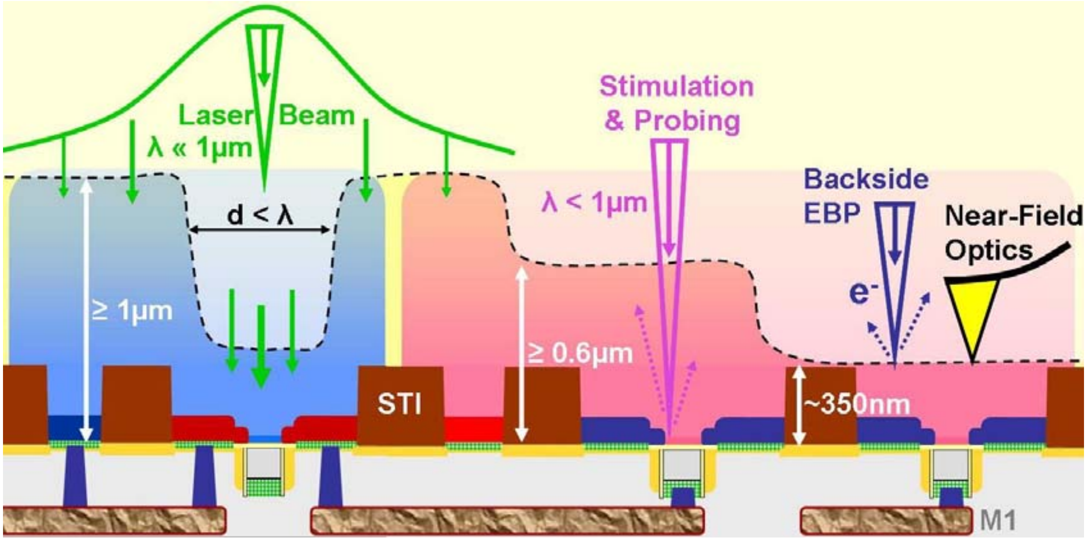

Figure 10-1: Optical and electron beam based probing and stimulation techniques on FIB thinned DuT

10.1. Laser-Optical Probing and Stimulation

Due to the absorption properties of silicon, all established optical backside probing and stimulation techniques are bound to use infrared laser light [7] (λ = 1064 and 1300 nm are the most commonly used wavelength). The theoretical resolution limit scales linearly with the used wavelength. Figure 10-2 shows the calculated minimum transmitted wavelength as a function of the remaining low doped Si thickness. Hereby it was assumed that the light has to pass the Si twice and that the overall absorption should not exceed 90% to allow for sufficient signal to noise ratio.

Substantially reducing the Si thickness in the area of interest allows decreasing the wavelength, increasing the resolution. Considering the results of the FIB invasiveness evaluation, the thickness of approximately 1 μm (green area in Figure 10-2) seems to be the most suited, since it allow for approximately twice the resolution (λ ≈ 0.6 μm) and still has no significant impact on device performance. But to gain a real resolution improvement compared to today's commercially available tools, additional SIL solutions based on materials being transparent to visible light would have to be established.

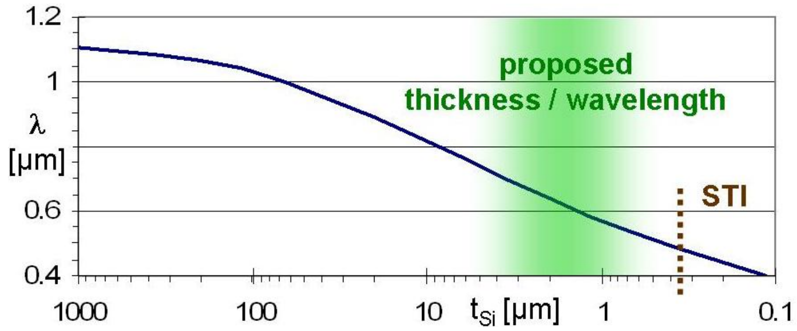

Figure 10-2: Calculated minimum wavelength with 10% transmission after passing the remaining Si (t_{Si}) twice (in- plus out-coming path), illustrating the potential resolution gain with working at reduced wavelength on thinner DuTs

Another approach is based on using the well-material as a kind of aperture. A possible scenario is illustrated on the very left of Figure 10-1. Again, the Si thickness is reduced to ≥ 1 µm in the vicinity of the targeted gates. In the next step, the FIB is utilized to mill a smaller access hole, reducing the thickness only in the area right above the targeted FET (or gate) to below 500 nm. Now, applying high energetic light (λ ≤ 400 nm), the whole incident beam energy is absorbed in the uppermost 200 nm. Besides the hole, all generated electron hole pairs can dissipate via the low ohmic well-contacts and only the targeted FET is stimulated. Regarding the geometrical properties of the proposed FIB milled light access hole, the upper diameter would have to be ≥ λ/2, to allow enough light-energy to be fed into the hole. Furthermore, the aspect ratio of such holes would have to be kept very low (< 5) to ensure that a realistic incident beam energy density yields sufficient device performance stimulation. The major benefit of this approach is that the well-Si aperture would decouple the beam diameter, or optical resolution from the achievable lateral resolution considering the desired localization and confinement of the laser stimulation. Until now, this is only a concept and will have to be evaluated in future projects.

10.2. Near-Field Optics on UtS

The concept of scanning optical near field microscopy (SNOM) has already proven its capability of providing nanoscale resolution independent of the used wavelength [33]. But the successful application requires near field conditions between the tip and the desired region of interaction, hence a working distance far less than the used wavelength. Until now, this limited the applicability of the technique for IC FA to "proof of concept" like lap conditions, where isolated test FETs in the µm scale were investigated through the unstructured frontside. On real chips, this access would be blocked by the backend.

Applying this technique to the rear surface of extensively FIB thinned samples (UTS) may allow expanding the applicability of SNOM to up to date VLSI DuTs. Investigations have just begun.

10.3. Backside Electron Beam Probing

The most promising of the FIB thinning (UtS) related circuit analysis approaches is backside electron beam probing (EBP). EBP on front side devices has been very useful for over two decades [34] and the underlying principal is depicted in the left half of Figure 10-3. But the capacitive coupled voltage contrast (CCVC) delivers only sufficient signal strength when measuring on the uppermost two metal layers. Consequently, this technique mostly fails for recent high metal stack technologies.

With the FIB thinning to STI level, the distance between M1 and rear surface is reduced to ≤ 1 µm which is sufficiently low to detect a CCVC signal, but with today's integration density 1 µm is by fare too much to separate the signals of adjacent M1 lines, hence cross talk makes this approach unpractical.

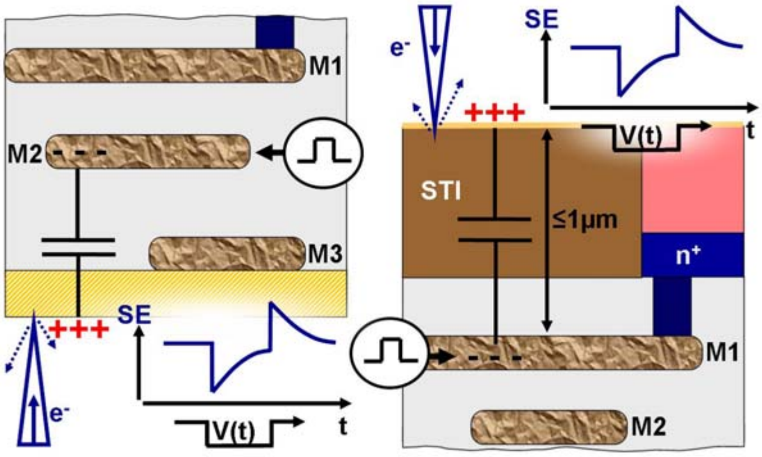

Figure 10-3: Comparison between the geometries of suitable frontside and backside EBP-samples when targeting metal lines via CCVC

Applying the proposed backside FIB thinning to CMOS gates of interest, the already discussed modulation of the floating body potential can now be interpreted as a signal, comprising the exact timing information of the targeted node which can be directly read out by EBP. Figure 10-4 depicts the geometry of a typical test structure. As seen in chapter 8, the whole body follows the drain potential, hence the full rear surface of the thinned FET carries the timing information, and the probe diameter (striped circle) may have the size of the full FET body, reducing the resolution requirements for the EBP. Another major advantage of this approach is its robustness against cross-talk. Considering signal 1 to be the target, any disturbing signal on an adjacent metal line (signal 2) can only weakly influence the surface potential above the non conductive STI (via C_y), but the conductive nature of the FET source-, drain- and gate-areas block any capacitive coupling into the FET body (C_x is uncritical). Furthermore, neighboring FETs will only have negligible influence (via C_z) once the devices are thinned to STI level which increases $R_{surface}$ into the GΩ regime.

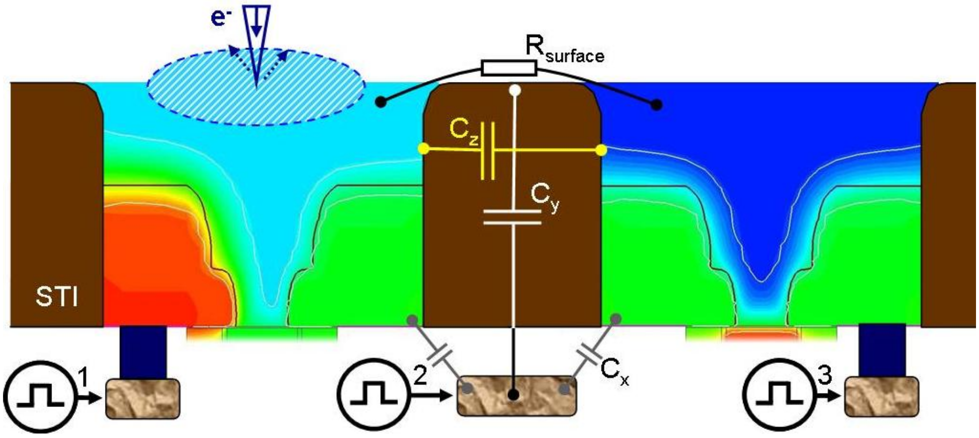

Figure 10-4: Relevant geometry parameters for the evaluation of cross-talk for backside EBP

The superior resolution potential of this technique was demonstrated on a 90 nm bulk technology test-chip by use of a > 13 year old EBP tool. Figure 10-5 shows a readily FIB prepared test location with the FIB image on the left, captured right after trenching down to STI level (with CAD layout overlay) whereas the right image was captured using the EBP in imaging mode. The DuT was powered up by $V_{dd} = 1.2$ V in the EBP, provoking a small static voltage contrast between the p- and n-FET body surface.

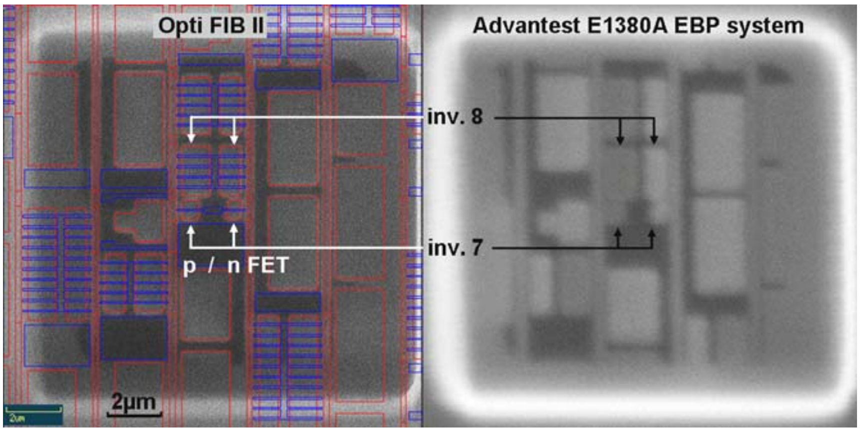

Figure 10-5: SE images of 90 nm DUT, captured in FIB (left) and EBP (right), with two subsequent inverters

Inverter 7 is one of the smallest gates of this process technology, could be resolved clearly and measured even without CAD overlay (n-FET area ≈410 x 690 nm^2). The results of the two subsequent inverters are shown in Figure 10-6.

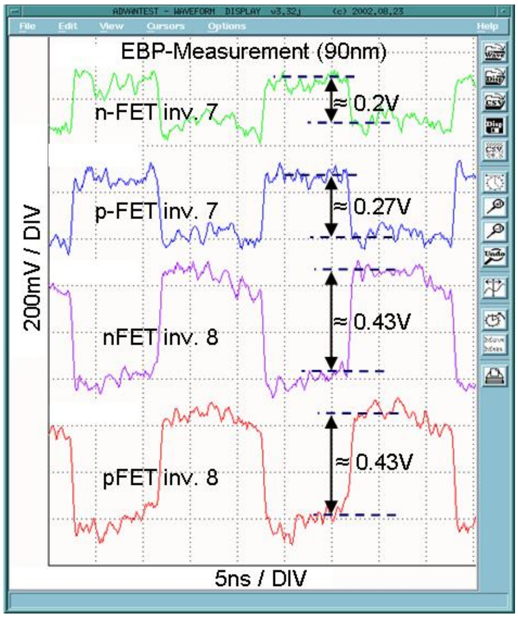

Figure 10-6: EBP measurement results on 90 nm technology, with 50 MHz input signal, 128 averages and 150 ps e-beam pulse width (probe point placement without CAD overlay)

Going from the first inverter 7 to the next one, the signal shows the expected 180° phase shift. Both gates are pure inverters (neither NAND, nor NOR based) and the modulation of the drain capacitance should result in an equivalently strong modulation of the body potential with all four measured FETs. Even though the probe diameter of the 1 keV electron beam is specified to be ≤ 100 nm [35], the reduced signal amplitude measured on inverter 7 may be due to the smaller geometries, where the pulsed e-beam diameter either already exceeds the active FET area or was not placed perfectly centric. Nevertheless, also with smallest geometries, the signal quality remains superior in terms of signal to noise ratio and especially according to acquisition time (≈60 s with all shown traces in Figure 10-6). Prior measurements on 180 nm inverter chains could already prove the ability of measuring arbitrary signals and the linear dependency between signal amplitude and V_{DD} (Figure 10-7), allowing the application of this technique at very low voltages, where acquisition times of optical techniques like time resolved emission (TRE) are already excessively long. Furthermore, the amplitude level and dependency supports the simulation based explanation model of the FIB induced speed gain discussed in chapter 7.

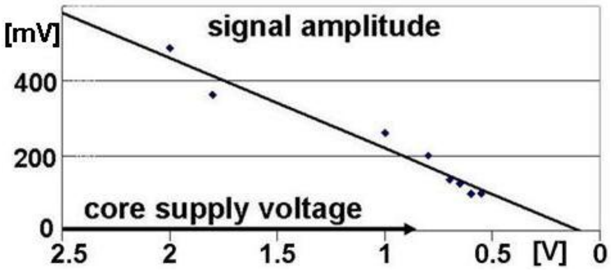

Figure 10-7: EBP signal amplitude vs. core supply voltage, measured at 180 nm inverter chains, similar to the test structures shown

The main drawback of backside EBP is the extensive preparation effort. Compared to the established optical techniques, here every measurement point (MP) needs to be pre-planned and FIB processed in advance to the measurement. The biggest overhead is due to the mechanical thinning (≥ 3 h, greatly depending on the package, tools, experience level…) followed by the coarse FIB thinning to n-well (≥ 1 h in FIB, depending on the quality of the mech. prep., see chapter 3.2 for more details). With the readily prepared trench to n-well it takes less than two minutes to open a MP. Considering today’s chip density, whole circuit blocks fit into one n-well trench area. Consequently, with the ongoing technology shrinkage the probability rises that all necessary MPs can be placed in one n-well trench area, which further decreases the preparation overhead.

Considering a typical backside CE, the readily prepared trench can be used for either CE verification or additional probing task without any additional effort. The situation is depicted in the right half of Figure 10-8, where the metal bridge of the CE (blue line) can be directly utilized to read out the node information (MP5).

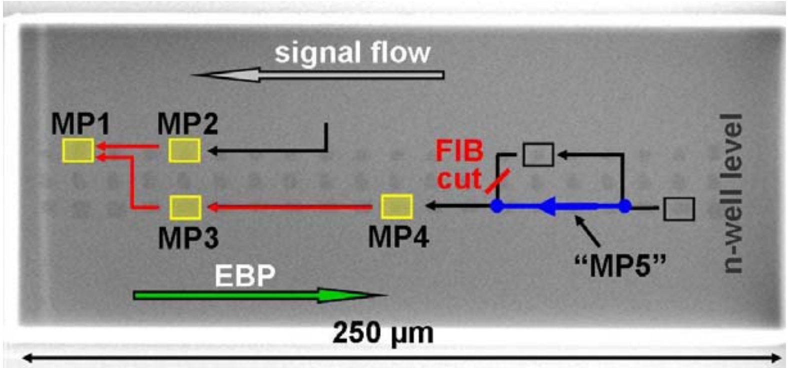

Figure 10-8: FIB image of trench to n-well level with illustration of a typical backside CE (red cut & green reconnection), including the proposed order of EBP MP

Regarding the accuracy of the measured timing information, the impact of the FIB thinning procedure has to be taken into account. The speed enhancement of a thinned gate only accelerates the gate output and does not affect the incoming signal or any gate before. Consequently, adulterating influences can be minimized by following the natural way of soft-fail localization, tracking the timing problem from signal output towards input. In the depicted example the binary search would start with MP1 and continue subsequently against the signal flow with MP2 and MP3 in the second step and so on.

The second severe problem is the intensified self-heating due to the extensive mechanical and FIB thinning. DuTs with high power density require adequate in vacuum cooling solutions which still have to be developed.

The biggest advantage of backside EBP is its resolution potential in combination with the extremely fast acquisition times. The available (≥ 13 year old) EBP tools have specified minimum spot sizes of less than 100 nm (e.g. [35] & [36]), which is already better than what current optical tools incorporating SIL technology can offer. Regarding that the EPB signal is present on the full FET body ($\geq 3 \cdot L_G$) even the smallest transistors in the upcoming 30 nm technology node should still be suitable measurement targets. Considering the progress of modern low keV SEM tools in the past decade (e.g. [37]) and the recent development of laser pulsed electron sources [38], pulsed spot sizes below 10 nm with sufficient timing resolution for upcoming technology generations seem to be possible, making backside EBP one of the most promising candidates for next generation high resolution probing.

A detailed comparison between backside EBP and the established optical probing techniques was published in [39].

11. Summary and Outlook

Fast and reliable failure analysis and design debug is critical for speeding up the development of new semiconductor products, and to increase profit due to reduced time to market. But, with the ongoing technology development (e.g. reduction of minimum feature sizes, package innovations, new front- and back-end materials….) all employed circuit analysis and modification techniques are continuously challenged to keep pace with the progress of the technologies to be worked on.

The key challenge for backside circuit edit (CE), in accordance to the technology development, originates from the ever decreasing structure sizes, requiring better ion-beam placement accuracy (alignment) and tool stability. The main goal of this work was to investigate the invasiveness, and to further improve the robustness and applicability of the proposed backside CE process, bearing up to these requirements.

The proposed procedure uses the STI-alignment, where parts of the DuT are ultimately thinned down to the bottom level of the 350 nm tall STI shapes. The respective thickness level is referred to as ultra thin silicon (UtS). This procedure has proven to provide highest alignment accuracy, but at the cost of altering the performance of the underlying transistors.

Considering backside CE, only small areas (2-4 µm^2) are thinned to UtS, having uncritical influence on the circuitry in most cases. The high local alignment accuracy, combined with the reproducible geometrical conditions of UtS allowed developing new contact methodologies, as a part of this work. The most promising approach is the so called contact to Silicide (CtS), where the node access hole is milled through the transistor body (as far apart from the gate as possible) and stopped in the Silicide layer, which builds the interface between the tungsten contacts (backend) and the highly doped drain/source diffusion areas (frontend). The methodology has shown reproducible and low ohmic contact properties, was successfully applied for circuit edit, direct probing and backside device characterization, and can help to sustain the applicability of backside circuit edit for future technology generations.

The FIB-thinned rear surface can also function as a platform for circuit analysis with increased resolution. The established optical circuit analysis techniques can benefit, since the reduced bulk Si thickness allows using shorter wavelength, hence increasing the resolution. Several different approaches have been discussed and can be subject of future investigations.

The application of electron beam probing on FIB-thinned transistors could be successfully demonstrated as a part of this work. Measurement results, captured on 90 nm technology devices, have shown superior lateral resolution and acquisition speed, compared to the established optical tools. The specified lateral resolution of the available EBP systems (spot-sizes < 100 nm, tools older than 13 years) should be sufficient to measure even smallest transistors in the upcoming 30 nm technology, making this technique a promising candidate for high resolution probing in the near future.

But, creating UtS in wider areas, as necessary for the above described analysis purposes, requires precise knowledge about the FIB induced device performance alterations. For this reason, an in-depth invasiveness study of the proposed UtS fabrication was carried out on single FETs and differently designed ring oscillators (ROs), all fabricated in 120 nm bulk technology, building the centre part of this work.

The DC-characterized single FETs suffer from increased self-heating and channel mobility degradation, reducing their performance by ≈ 10% on UtS level. In addition, the initially low ohmic connection between the FET body and the covering/surrounding well-material is turned high ohmic, finally being almost fully disconnected, inducing SOI-related phenomena like the Kink-effect.

In contrary to the DC performance degradation, the FIB thinning increases the circuitry performance by 10% up to 60%, when running a CMOS chip in dynamic operation. The speed gain originates from a reduction of the FET-intrinsic drain junction capacitance (similar to SOI-FETs) in combination with a reduction of the threshold voltage. The capacitance reduction is design/layout dependent, hence easily predictable. The performance increase resulting from the V_t reduction originates from an increased FET body potential, which is strongly dependent on the remaining device thickness, the used supply voltage level and several technology parameters (e.g. well-design, source/drain implant profiles….) and comes with a higher forecast complexity.

Connecting and complementing the experimental data with physical device simulation results (using the Synopsys® Sentaurus numerical simulation environment) allowed deriving the first ever reported quantitative degradation model, valid for the static and dynamic device operation. Based on these results, it is now possible to predict the behavior of different CMOS designs, fabricated in different bulk technologies, being subject to FIB-thinning.

The demonstrated, and now also predictable FIB induced speed gain can therefore be used to cope with another increasingly problematic bottleneck in the chip development, related FA and debugging, which is dealing with soft-fails.

Chip-internal timing misalignments are already the predominant limitations of high speed designs, and the often very small deviations from the simulated circuit model are not only hard to pin-point but also almost impossible to correct by means of classical circuit edit. Early chip designs often include additionally implemented structures (e.g. varactors and fuses) to allow for timing adjustments in the readily build chips, but these solutions are cost-intensive (area overhead) and consequently often not there when needed.

The FIB-thinning induced speed gain can now be utilized for trimming the chip-internal timing conditions. The proposed procedure can be applied to any single gate on a give chip, delivering from 10% up to 60% performance increase, depending on the various identified parameters (layout/design, V_{dd}, t_{Si}…). The high level of process control and the unsurpassed trench-floor co-planarity allows the secure and reproducible UtS formation in chip areas wider than 100x100 μm^2, speeding up full circuit block with one FIB trench. The FIB thinned DuTs remain fully functional, which may allow for the production of engineering samples, being critical for early prototyping and customer pre-qualification, which is often impossible only utilizing the currently established methodologies.

In general, the FIB-created ultra thin Silicon can become a new platform for circuit edit, circuit analysis, stimulation and probing, offering increases resolution, and the same approach allows for trimming of chip internal timing conditions.

12. References

[1] J. Melngailis, et al., “The focused ion beam as an integrated circuit restructuring tool”, Journal of Vacuum Science & Technology B, Volume 4, Issue 1, 1986, pp.176-180

[2] K. (Bobby) Hooghan, “Focused Ion Beam (FIB) Systems: A Brief Overview”, Microelectronics Failure Analysis: Desk Reference Fifth Edition, 2004, pp. 583-594

[3] R. J. Young, et. al., “Focused ion beam insulator deposition”, Journal of Vacuum Science & Technology B, Volume 13, Issue 6, 1995, pp. 2576 - 2579

[4] I. Utke, et. al.,” Gas-assisted focused electron beam and ion beam processing and fabrication”, Journal of Vacuum Science & Technology B, Volume 26, Issue 4, 2008, pp. 1197 - 1276

[5] S.X. Li, et al., “Performing Circuit Modification and Debugging Using Focused-Ion-Beam on Multi-Layered C4 Flip-Chip Devices”, Proc. ISTFA 1998, pp. 67-72

[6] U. Kindereit et al., “Investigation of Laser Voltage Probing Signals in CMOS Transistors”, Proc. IEEE IRPS 2007, pp 526-533

[7] J.C.H. Phang, et al., “A review of laser induced techniques for microelectronic failure analysis”, Proc. IEEE IPFA 2004, pp 255-261

[8] Mario Pannicia et al., “Novel Optical Probing Technique for Flip Chip Packaged Microprocessors”, Proc. IEEE ITC 1998, pp.740

[9] R.J. Livengood, et al., ”Application of advanced micromachining techniques for the characterization and debug of high performance microprocessors”, Journal of Vacuum Science & Technology B, Volume 17, Issue 1, 1999, pp. 40-43

[10] C. Boit, et al., ”Impact of Backside Circuit Edit on Active Device Performance in Bulk Silicon ICs”, Proc. IEEE ITC 2005, pp. 48.2

[11] M. Abramo, et al., “FIB Backside Isolation Techniques”, Microelectronic Failure Analysis: Desk Reference 2001 Supplement, pp. 1-17

[12] C. Boit, et al., “Physical IC debug – backside approach and nanoscale challenge”, Advances in Radio Science, 2008, Issue 6, pp. 265–272

[13] K. Ng, et al., “FIB etching of Cu with minimal impact on neighboring circuitry, including dielectric”, Proc. IEEE IPFA 2005, pp 118-122

[14] H. Gnaser, et al., “Focused ion beam implantation of Ga in Si and Ge: fluence-dependent retention and surface morphology”, Surface and Interface Analysis, 2008, Volume 40, Issue 11, pp. 1415-1422

[15] Q. Gao, et al., “Experiment study on Crystal/Amorphous Structure of TEM Samples Prepared by FIB Milling”, Proc. ISTFA 2006, pp. 76-79

[16] Q. Wang, et al., “Modeling Secondary Electron Emission from High Aspect Ratio Holes”, Proc. ISTFA 2003, pp. 343

[17] V. Mararov, et al., “Milling High Aspect Ratio (HAR) Holes in Dielectrics”, European FIB user-group meeting (EFUG) 2005, Arcachon, France

[18] F. Mosselveld, et al., "Circuit editing of copper and low-k dielectrics in nanotechnology devices", Journal of Microscopy, 2004, Vol. 214, Pt 3, pp. 246–251

[19] M.A. Thompson, et al., "Coaxial, Photon-Ion Technology Enables Direct Navigation to Buried Nodes on Planarized Surfaces, including Silicon", Proc. 28th ISTFA 2002, pp. 409

[20] Yuan Taur, Tak H. Ning, "Fundamentals of Modern VLSI Devices" Cambridge University Press, 1998

[21] J. Gautier, et al. "Body Charge Related Transient Effects in Floating Body SOI NMOSFET´s", Proc. IEEE IEDM 1995, pp. 623

[22] A. Wei, et al. "Effect of Floating Body Charge on SOI MOSFET Design", Transactions on Electron Devices, 1998, Vol. 45, No. 2, pp. 430

[23] Synopsys® Sentaurus Structure Editor User Guide, Version Z-2007.03, March 2007

[24] Synopsys® Sentaurus Device User Guide, Version Z-2007.03, March 2007

[25] D. B. M. Klaassen, "A Unified Mobility Model for Device Simulation - I. Model Equations and Concentration Dependence", Solid-State Electronics, 1992, vol. 35, no. 7, pp. 953–959

[26] M. N. Darwish et al., "An Improved Electron and Hole Mobility Model for General Purpose Device Simulation", IEEE Transactions on Electron Devices, 1997, vol. 44, no. 9, pp. 1529–1538

[27] C. Canali et al., "Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature", IEEE Transactions on Electron Devices, 1975, vol. ED-22, no. 11, pp. 1045–1047

[28] Keithley 4200-SCS, Semiconductor Characterization System, Reference Manual, by Keithley 2007

[29] R. Schlangen, et al. "Trimming of IC Timing and Delay by Backside FIB Processing - Comparison of Conventional and Strained Technologies", Proc. IEEE IEDM 2008, pp. 439-442

[30] R. Schlangen, et al., "Physical Analysis, Trimming and Editing of Nanoscale IC Function with Backside FIB Processing" Microelectronics Reliability, 2009, Volume 49, Issues 9-11, pp. 1158-1164

[31] R.K. Jain et al., "Advanced Fringe Analysis Techniques in Circuit Edit", Proc. ISTFA 2006, pp. 79-85

[32] R. Schlangen, et al.,"New Circuit Edit and Probing Options directly to FET Device on Ultra Thin Silicon Backside processed by Focused Ion Beam", IEEE-SDD workshop, 2007, Freiburg, Germany

[33] D.V. Isakov, et al., "Applications of Scanning Near-Field Photon Emission Microscopy", Proc. ISTFA 2008, pp. 25-29

[34] W.T. Lee, et al., "Engineering a Device for Electron Beam Probing", IEEE Design and Test of Computers, June 1989, pp. 36

[35] Datasheet: E1380A, part of: E1380A EB Test System Main Frame Operation Manual, by Advantest 1999

[36] Datasheet: IDS 10000$_{da}$, by Schlumberger 1999

[37] J.P. Vermeulen, "New Developments in FESEM Technology", Advanced Materials & Processes, 2005, Volume 163, Issue 8, pp. 33

[38] M. Merano, et al., "High brightness picosecond electron gun", AIP - Review of Scientific Instruments 76, 085108 (2005); doi:10.1063/1.2008975

[39] R. Schlangen, R. Leihkauf, U. Kerst, C. Boit, K. Wilsher, T. Lundquist et al., "Backside E-Beam Probing on Nano Scale Devices", Proc. IEEE ITC 2007, pp. 23.2 1-9

13. Chronological list of Publications

I. R. Schlangen, U. Kerst, A. Kabakow, C. Boit "Electrical Performance Evaluation of FIB Edited Circuits through Chip Backside Exposing Shallow Trench Isolations", Proc. ESREF 2005, pp. 1544

II. C. Boit, U. Kerst, R. Schlangen, A. Kabakow, E. Le Roy, T.R. Lundquist, S. Pauthner "Impact of Backside Circuit Edit on Active Device Performance in Bulk Silicon ICs", Proc. IEEE ITC 2005, pp. 48.2

III. R. Schlangen, P. Sadewater, U. Kerst, C. Boit "Contact to Contacts or Silicide by use of Backside FIB Circuit Edit allowing to approach every Active Circuit Node", Proc. ESREF 2006, pp. 1498

IV. R. Schlangen, R. Leihkauf, U. Kerst, C. Boit, B. Krüger "Functional Analysis Through Chip Backside with Nano Scale Resolution", Proc. ISTFA 2006, pp. 376
Best Paper Award

V. R. Schlangen, U. Kerst, C. Boit, R. Jain, T. Malik, T. Lundquist "New Circuit Edit and Probing Options directly to FET Device on Ultra Thin Silicon Backside processed by Focused Ion Beam", IEEE-SDD workshop, 2007, Freiburg, Germany

VI. R. Jain, T. Malik, T. Lundquist, R. Schlangen, R. Leihkauf, U. Kerst, C. Boit "Novel Flip-Chip Probing Methodology using Electron Beam Probing", Proc. IEEE IPFA 2007, pp. 39 Best Paper Award

VII. R. Schlangen, U. Kerst, C. Boit, R. Jain, T. Malik, T. Lundquist "Non Destructive 3D Chip Inspection with Nano Scale Potential by use of Backside FIB and Backscattered Electron Microscopy", Proc. ESREF 2007, pp. 1523

VIII. R. Schlangen, R. Leihkauf, U. Kerst, C. Boit, R. Jain, T. Malik, K. Wilsher, T. Lundquist, B. Krüger "Backside E-Beam Probing on Nano Scale Devices", Proc. ITC 2007, pp. 1

IX. R. Schlangen, U. Kerst, C. Boit, S. Schömann, B. Krüger, R. Jain, T. Malik, T. Lundquist "FIB backside circuit modification at the device level, allowing access to every circuit node with minimum impact on device performance by use of Atomic Force Probing", Proc. ISTFA 2007, pp. 34

X. C. Boit, R. Schlangen, U. Kerst, T. Lundquist "Physical Techniques for Chip-Backside IC Debug in Nanotechnologies", IEEE Design & Test of Computers, Special Issue: Silicon Debug and Diagnosis, 2008, pp. 250

XI. R. Schlangen, R. Leihkauf, T. Lundquist, P. Egger, U. Kerst, C. Boit "Trimming of IC Timing and Delay by Backside FIB Processing - Comparison of Conventional and Strained technologies", Proc. IEEE IEDM 2008, pp. 439

XII. R. Schlangen, R. Leihkauf, T. Lundquist, P. Egger, C. Boit "RF Performance Increase Allowing IC Timing Adjustments by Use of Backside FIB Processing", Proc. IEEE IPFA 2009, pp. 33

Invited Speaker at ESREF 2009

XIII. R. Schlangen, R. Leihkauf, T. Lundquist, P. Egger, C. Boit "Physical Analysis, Trimming and Editing of Nanoscale IC Function with Backside FIB Processing", Microelectronics Reliability, 2009, Volume 49, Issues 9-11, pp. 1158-1164

XIV. R. Schlangen, R. Leihkauf, U. Kerst, C. Boit, P. Egger, T. Lundquist "Extended Circuit Edit, Analysis and Trimming Capabilities based on the Backside Focused Ion Beam created Ultra Thin Silicon Platform", to be published at ISTFA November 2009

14. Acknowledgments

The author would like to thank the following individuals for their support during the development of this work.

Special thanks go to Prof. Dr.-Ing. Christian Boit for guiding me in his disarmingly-positive way, for being a sheer endless source of new ideas, and for granting me the freedom to “play” in the Lab, following up on the ideas I liked most. I hope you did not regret hiring your first “blind-date” PhD-student.

I wish to thank Prof. John Melngailis for his interest in my work and the fruitful discussions. It was a real pleasure and honor to get to know you, one of the FIB pioneers.

Another important person throughout the past fife years was Dr. Theodore R. Lundquist, who strongly supported my work, has always been a great host in the US and around the globe, and helped me to understand many important things which are not thought at Universities.

Furthermore in want to thank the other people at DCG-Systems who took good care of me, namely: Rajesh Jain, Tahir Malik, Surendra Madala, Ketan Shah, Mark Thompson, Ken Wilsher, Vladimir Makarov, Marie Headley and many more. See you soon!

Next: Rainer Leihkauf, who spend endless hours in front of the simulations and helped me a lot in understanding the results. Thank you so much for your help, I wish you all the best!

Other very important colleges were: Andreas Eckert, who helped me from the first day on and prepared the thinnest and flattest backside samples one can imagine, Uwe Kerst, for patiently teaching me FIB and helping with many publications, Philipp Scholz, for helping me so much with teaching and for being a nice guy and very good office-neighbor, and Tuba, Piotr, Arek, Taro, Mahyar, Ulrike, Peter, Silvia, Marlies, Helmut for their help and the positive atmosphere at work.

I would also like to thank Dr. Siegfried Görlich, Peter Egger, Bernd Krüger, Stephan Schömann, Christof Brillert and Siegfried Pauthner of the Infineon FA team in Munich for their continuous support of my research activities and especially for the opportunity to spend time in their extremely well equipped and well organized Laboratory.

Letztlich, aber sicherlich nicht zuletzt, möchte ich meiner ganzen Familie, und ganz besonders meiner Mutter, danken, ohne deren Hilfe ich sicher nie eine Uni von innen gesehen hätte, und die mich (und ich Sie) leider die letzten Jahre viel zu wenig gesehen haben. Vielen Dank!

15. Appendix

(A) Vacuum Setup (A)

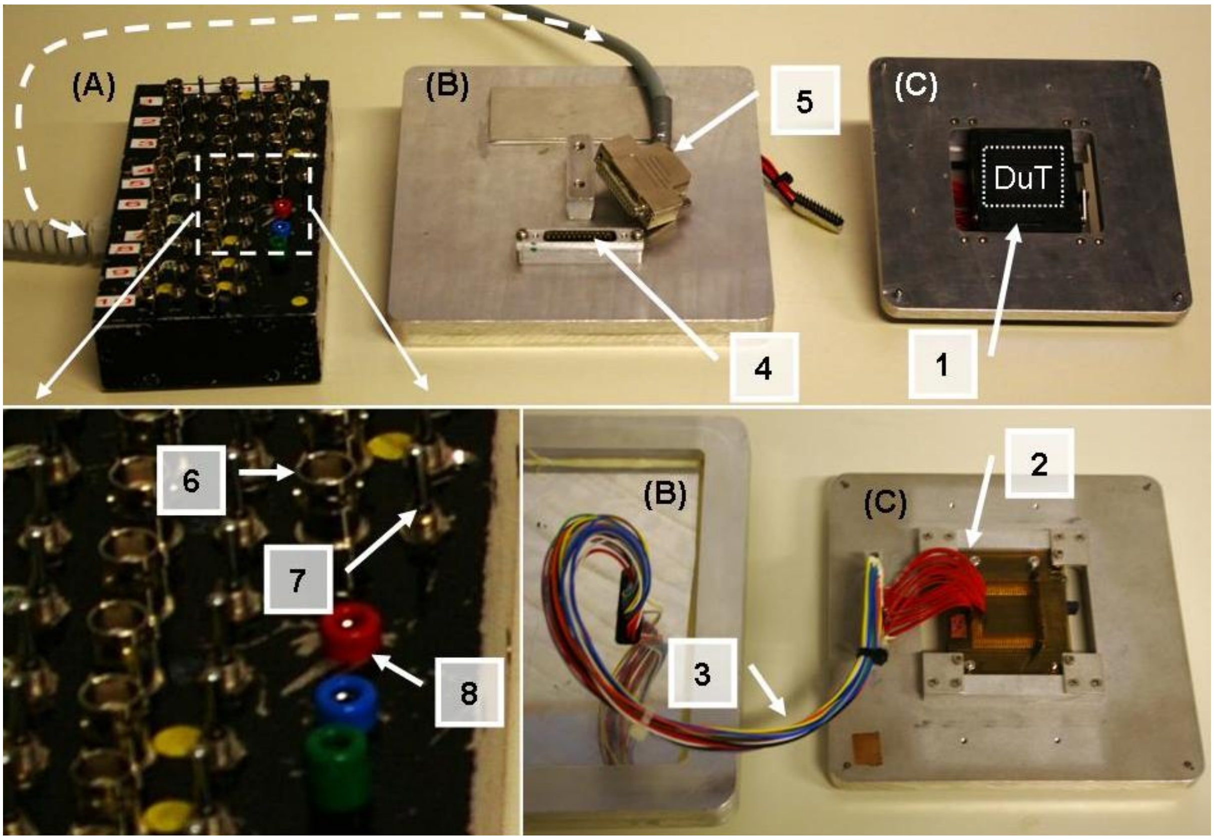

Figure 15-1: Vacuum Setup (A)

(A) BNC-Connection Box

(B) Lit of vacuum chamber with feed-through

(C) In-vacuum part, with SBGA socket

The setup consists of three parts, as listed above. The DuT is located in (C), in the OptiFIB vacuum chamber, is contacted via an SBGA burn-in socket (1) and can move in z-direction to adjust the optical focus in the FIB tool. Since the vacuum feed-through is limited to 25 pins, the connection to the rear PGA of the SGBA socket was made flexible (with 2 plugs, connecting once 13, once 12 neighboring pins (2)) to allow for easy setup modification. The 25x vacuum feed-through is mounted in the top-lit of the vacuum chamber (B) and 25 lose wires (3) connect it to (C). Outside of the FIB, a shielded standard 25x sub-D connector in combination with a shielded 25x data cable (5) is used to link the 25x BNC Box (A) with the DuT. The BNC Box is equipped with 25 standard BNC connectors, and each one has a three-step switch (7), allowing to apply DC voltages, which can be supplied via (8).

The utilized test-structures correspond to the following pin-lists as follows:

10x0.12 µm n-FET → n FET 6, 10x0.12 µm p-FET → p FET 6

RO (A) → R3, RO (B) → R4, RO (C) → R5

Pin-list for utilized 120 nm samples (n-FETs & ROs)

AA	AB	AC	AD	AE	AF		SBGA pins
			plug (2) position for ROs		Gnd-loop	1	
						2	
				1		3	
				2	En R6	4	
			14	3	En R5	5	
		Out R6	15	4	En R4	6	
		Out R5	16	5	Vss R4-6	7	
		Out R4	17	6	Out R3	8	
	plug (2) position for n-FETs		18	7	En R3	9	
		Vdd R4-6	19	8	En R2	10	plug (2) position for n-FETs
	14	Out R2	20	9	Vdd R2-3	11	
	15	Vss R2-3	21	10	D nFET 6	12	1
	16	Gnd-loop	22	11		13	2
	17	S nFET 6	23	12		14	3
	18	S nFET 5	24	13	G nFET 6	15	4
	19	D nFET 4	25		D nFET 5	16	5
	20	D nFET 3			G nFET 5	17	6
	21				S nFET 4	18	7
	22	D nFET 2			G nFET 4	19	8
	23	G nFET 2			S nFET 3	20	9
	24	S nFET 1			G nFET 3	21	10
	25	p-well			S nFET 2	22	11
					D nFET 1	23	12
					G nFET 1	24	13
						25	
						26	

Pin-list for utilized 120 nm samples (p-FETs & non-standard ROs)

F	E	D	C	B	A	SBGA pins	
			plug (2) position for p-FETs		Gnd-loop	26	
						25	
				1	n-well	24	
				2		23	
		p-well	14	3	S pFET 2	22	
			15	4	G pFET 3	21	
		G pFET 2	16	5	S pFET 3	20	
		D pFET 2	17	6	G pFET 4	19	
	plug (2) position for non-stan. ROs		18	7	S pFET 4	18	
		D pFET 3	19	8	G pFET 5	17	plug (2) position for non-stan. ROs
	14	D pFET 4	20	9	D pFET 5	16	
	15	S pFET 5	21	10	G pFET 6	15	1
	16	Gnd-loop	22	11		14	2
	17	S pFET 6	23	12		13	3
	18		24	13	D pFET 6	12	4
	19	Vss R2-3	25			11	5
	20	En R3			Vdd R2-3	10	6
	21				En R2	9	7
	22	Vss R4-6			Out R2	8	8
	23	Out R4			Out R3	7	9
	24	Out R5			Vdd R4-6	6	10
	25	Out R6			En R4	5	11
					En R5	4	12
					En R6	3	13
						2	
						1	

(B) Synopsys Sentaurus definition -files

(B1) n-FET Device Definition file (t_{Si} = 480 nm):

```
(if (string=? "nMOS" "nMOS")

;nMOS,

 (begin

; structural parameters  SPACER

(define Lsp1   0.007)
(define Lsp2   0.06 )

; Doping Implants

; Substrate
(define DopSub "BoronActiveConcentration")
(define SubDop  1e16 )

; Poly doping
(define DopPoly
"PhosphorusActiveConcentration")
(define PolyDop  6e18 )

; SD 1
(define DopSD "PhosphorusActiveConcentration")
(define SDDop    2e20 )
(define XpSD     0.02 )
(define XjSD     0.057)
(define VaXjSD   1e19)

; SD 2
(define DopSD2
"PhosphorusActiveConcentration")
(define SDDop2    1e19 )
(define XpSD2     0.077)
(define XjSD2     0.133)
(define VaXjSD2   1e17)

; Vt doping
(define DopVt "BoronActiveConcentration")
(define VtDop  2.5e17 )
(define XpVt   0.105)
(define XjVt   0.135)
(define VaXjVt   8e16)

; Vt2 doping
(define DopVt2 "BoronActiveConcentration")
(define VtDop2    100)
(define XpVt2   0.105)
(define XjVt2   0.135)
(define VaXjVt2     1)

;Ret dop A
(define DopRetA "BoronActiveConcentration")
(define RetDopA    4e18 )
(define XpRetA    0.8 )
(define XjRetA    0.42  )
(define VaXjRetA  1e17  )

;Ret dop B
(define DopRetB "BoronActiveConcentration")
(define RetDopB  3e17 )
(define XpRetB   0.37 )
(define XjRetB   0.13 )
(define VaXjRetB 8e16 )

; Halo Dop
(define DopHalo "BoronActiveConcentration")
(define HaloDop        2e18)
(define XpHalo         0.015)
(define XjHalo         0.06)
(define VaXjHalo       5e17)
(define LatFacHelo    0.765)

; Ldd  SD Extention doping
(define DopExt "ArsenicActiveConcentration")
(define ExtDop          1e20 )
(define XpExt           0.017)
(define XjExt           0.065)
(define VaXjExt         5e17 )
(define LatFacLdd       0.37 )
 )
; end nMOS

; Setting common parameters
(define Tox    0.0022)
(define Hpol     0.2)
(define Lg   ( 0.12 0.015))
(define Xg   (/ Lg   2.0))
(define Lreox 0.007)
(define Xmax (+ (/ 0.12 2) 1.8 1.5))
(define Ysub   0.48)
(define STIdepth 0.35)

(if (< Ysub STIdepth)
 (begin
  (define STIde Ysub)
 )
 (begin
  (define STIde STIdepth)
 )
)
;    spacer und STI rounding
(define filletradius 0.05) ; [um] Rounding radius
```

```
; Derived quantities
(define Xsp1  (+ Xg   Lsp1))
(define Xsp2  (+ Xg   Lsp2))
(define Xrox (+ Xg   Lreox))
(define Ygox (* Tox 1.0))
(define Ypol ( Ygox Hpol))
(define Gpn 0.004)

; Overlap resolution: New replaces Old
(sdegeo:setdefaultboolean "ABA")

; Creating substrate region
(sdegeo:createrectangle
  (position    0.0  Ysub  0.0 )
  (position    Xmax 0.0 0.0 ) "Silicon"
"R.Substrate" )
  ; Creating gate oxide
(sdegeo:createrectangle
  (position    0.0 0.0  0.0 )
  (position    Xsp2 Ygox 0.0 )
  "SiO2" "R.Gateox"
)
; Creating spacers regions
(sdegeo:createrectangle
  (position    0.0   Ygox 0.0 )
  (position    Xsp2   Ypol 0.0 )
  "Si3N4" "R.Spacer"
)
; Creating PolyReox
(sdegeo:createrectangle
  (position    0.0     Ygox 0.0 )
  (position    Xrox  Ypol 0.0 )
  "Oxide" "R.PolyReox"
)
; Creating PolySi gate
(sdegeo:createrectangle
  (position    0.0 Ygox 0.0 )
  (position    Xg  Ypol 0.0 )
  "PolySi" "R.Polygate"
)
; Creating STI new
(sdegeo:createrectangle
  (position    ( Xmax 1.5)   0.0  0.0 )
  (position         Xmax      STIde 0.0 )
  "SiO2" "R.STI"
)
;  rounding spacer
(sdegeo:fillet2d
 (findvertexid (position Xsp2 Ypol 0.0 ))
 filletradius)
;  rounding STI left end
(sdegeo:fillet2d
 (findvertexid (position ( Xmax 1.5)  STIde  0.0 ))
 filletradius)

; Contact declarations
(sdegeo:definecontactset "drain"
  4.0  (color:rgb 0.0 1.0 0.0 ) "##")
(sdegeo:definecontactset "gate"
  4.0  (color:rgb 0.0 0.0 1.0 ) "##")
(sdegeo:definecontactset "substrate"
  4.0  (color:rgb 0.0 1.0 1.0 ) "##")

; Contact settings
(sdegeo:define2dcontact
 (findedgeid (position  (+ Xg Lsp2 1) 0.0 0.0))
 "drain")
(sdegeo:define2dcontact
 (findedgeid (position 5e4 Ypol 0.0))
 "gate")
(sdegeo:define2dcontact
 (findedgeid (position   Xmax ( Ysub 0.001) 0.0))
 "substrate")

; Separating lumps
(sde:separatelumps)

; Setting region names
(sde:addmaterial
  (findbodyid (position  5e4  ( Ysub 0.1) 0.0))
  "Silicon" "R.Substrate")
; gate Ox
(sde:addmaterial
  (findbodyid (position  5e4  (* 0.5 Ygox) 0.0))
  "SiO2"    "R.Gateox")
; poly gate
(sde:addmaterial
  (findbodyid (position  5e4  ( Ygox 0.01)  0.0))
  "PolySi"  "R.Polygate")
; nitride spacer right
(sde:addmaterial
  (findbodyid (position (*  0.5 (+ Xsp2 Xg)) ( Ygox
0.01)  0.0))
"Si3N4"   "R.Spacerright")
; STI right
(sde:addmaterial
  (findbodyid (position ( Xmax 0.4) (* STIde 0.5)
0.0))
"SiO2"   "R.STIright")

; Saving BND file
(sdeio:savetdrbnd (getbodylist)
"n475_half_bnd.tdr")

; Profiles:

;  Substrate
(sdedr:defineconstantprofile "Const.Substrate"
 DopSub SubDop )
(sdedr:defineconstantprofilematerial
"PlaceCD.Substrate"
 "Const.Substrate" "Silicon" )

;  Source/Drain implants 1
;  base line definitions
(sdedr:definerefinementwindow "BaseLine.SD"
"Line"
 (position     Xsp2         XpSD   0.0)
 (position  ( Xmax 1.5)    XpSD   0.0) )
;  implant definition auf 2e20
(sdedr:definegaussianprofile "Impl.SDprof"
 DopSD
 "PeakPos" 0  "PeakVal" SDDop
 "ValueAtDepth" VaXjSD "Depth" XjSD
 "Gauss"  "Factor" 0.3
)
;  implant placement
(sdedr:defineanalyticalprofileplacement "Impl.SD"
 "Impl.SDprof" "BaseLine.SD" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

;  Source/Drain implants 2
;  base line definitions
(sdedr:definerefinementwindow "BaseLine.SD2"
"Line"
 (position      Xsp2        XpSD2   0.0)
 (position  ( Xmax 1.5)    XpSD2   0.0) )
;  implant definition auf 2e20
(sdedr:definegaussianprofile "Impl.SD2prof"
 DopSD2
 "PeakPos" 0  "PeakVal" SDDop2
 "ValueAtDepth" VaXjSD2 "Depth" XjSD2
 "Gauss"  "Factor" 0.3
)
;  implant placement
(sdedr:defineanalyticalprofileplacement
"Impl.SD2"
 "Impl.SD2prof" "BaseLine.SD2" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

;  retrograde pwell A
;  base line definitions  neu
(sdedr:definerefinementwindow
"BaseLine.retWellA" "Line"
 (position       0       XpRetA   0.0)
 (position     Xmax      XpRetA   0.0) );nur bis zum
STI !!!!!
;  implant definition
(sdedr:definegaussianprofile "Impl.retprofA"
 DopRetA
 "PeakPos" 0  "PeakVal" RetDopA
 "ValueAtDepth" VaXjRetA  "Depth" XjRetA
 "Gauss"  "Factor" 0.8
)
;  implant placement
(sdedr:defineanalyticalprofileplacement
"Impl.retWellA"
 "Impl.retprofA" "BaseLine.retWellA" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

;  retrograde pwell B
;  base line definitions  neu
(sdedr:definerefinementwindow
"BaseLine.retWellB" "Line"
 (position       0       XpRetB   0.0)
 (position      Xmax      XpRetB   0.0) );nur bis zum
STI !!!!!

;  implant definition
(sdedr:definegaussianprofile "Impl.retprofB"
 DopRetB
 "PeakPos" 0  "PeakVal" RetDopB
 "ValueAtDepth" VaXjRetB  "Depth" XjRetB
 "Gauss"  "Factor" 0.8
)
;  implant placement
(sdedr:defineanalyticalprofileplacement
"Impl.retWellB"
 "Impl.retprofB" "BaseLine.retWellB" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

; Vt implant
;  base line definitions
(sdedr:definerefinementwindow "BaseLine.Vt"
"Line"
 (position       0             XpVt   0.0)
 (position   ( Xmax 1.5)    XpVt   0.0) )
;  implant definition
(sdedr:definegaussianprofile "Impl.Vtprof"
 DopVt
 "PeakPos" 0  "PeakVal" VtDop
 "ValueAtDepth" VaXjVt  "Depth" XjVt
 "Gauss"  "Factor" 0.8
)
;  implant placement
(sdedr:defineanalyticalprofileplacement "Impl.Vt"
 "Impl.Vtprof" "BaseLine.Vt" "Both" "NoReplace"
"Eval" "Silicon" 0 "material")

; Vt2 implant
;  base line definitions
(sdedr:definerefinementwindow "BaseLine.Vt2"
"Line"
 (position       0             XpVt2   0.0)
 (position   ( Xmax 1.5)    XpVt2   0.0) )
;  implant definition
(sdedr:definegaussianprofile "Impl.Vtprof2"
 DopVt2
 "PeakPos" 0  "PeakVal" VtDop2
 "ValueAtDepth" VaXjVt2  "Depth" XjVt2
 "Gauss"  "Factor" 0.8
)
;  implant placement
(sdedr:defineanalyticalprofileplacement "Impl.Vt2"
 "Impl.Vtprof2" "BaseLine.Vt2" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

;  Source/Drain extensions
; base line definitions
(sdedr:definerefinementwindow "BaseLine.Ext"
"Line"
 (position   Xsp1       XpExt  0.0)
 (position ( Xmax 1.5)   XpExt  0.0) )
;  implant definitionneu 6e19 1e17
(sdedr:definegaussianprofile "Impl.Extprof"
 DopExt
 "PeakPos" 0  "PeakVal" ExtDop
 "ValueAtDepth" VaXjExt "Depth" XjExt
 "Gauss"  "Factor" LatFacLdd
)
; implant placement
(sdedr:defineanalyticalprofileplacement "Impl.Ext"
 "Impl.Extprof" "BaseLine.Ext" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

; Halo
; base line definitions
(sdedr:definerefinementwindow "BaseLine.Halo"
"Line"
 (position ( Xsp1 0.007)  XpHalo   0.0)
 (position ( Xmax 1.5)    XpHalo   0.0) )
; implant definition neu 1e17 0.1
(sdedr:definegaussianprofile "Impl.Haloprof"
 DopHalo
 "PeakPos" 0  "PeakVal" HaloDop
 "ValueAtDepth" VaXjHalo  "Depth" XjHalo
 "Gauss"  "Factor" LatFacHelo
)
; implant placement
(sdedr:defineanalyticalprofileplacement
"Impl.Halo"
 "Impl.Haloprof" "BaseLine.Halo" "Both"
"NoReplace" "Eval" "Silicon" 0 "material")

; Poly
(sdedr:defineconstantprofile "Const.Gate"
 DopPoly PolyDop )
(sdedr:defineconstantprofileregion "PlaceCD.Gate"
 "Const.Gate" "R.Polygate" )

; Meshing Strategy:

; Substrate below pn and STI  maximum spacing
(sdedr:definerefinementsize "Ref.Substrate"
  0.2      0.2
  0.1      0.075 )
(sdedr:definerefinementfunction "Ref.Substrate"
 "DopingConcentration" "MaxTransDiff" 1)
(sdedr:definerefinementregion "RefPlace.Substrate"
 "Ref.Substrate" "R.Substrate" )
; side contact region
(sdedr:definerefinementwindow "RWin.ActSide"
 "Rectangle"
 (position   ( Xmax 0.05)    0.325   0.0)
 (position     Xmax          Ysub   0.0 ))
```

```
 (sdedr:definerefinementsize "Ref.SiActSide"
  0.05      0.05
  0.003    0.015 )
(sdedr:definerefinementfunction "Ref.SiActSide"
 "DopingConcentration" "MaxTransDiff" 1)
(sdedr:definerefinementplacement
"RefPlace.SiActSide"
 "Ref.SiActSide" "RWin.ActSide" )
; Active
(sdedr:definerefinementwindow "RWin.Act"
 "Rectangle"
 (position  0.0          0.0    0.0)
 (position  ( Xmax 1.4)     0.4     0.0 ))
 (sdedr:definerefinementsize "Ref.SiAct"
  0.1          0.01
  0.05          Gpn )
(sdedr:definerefinementfunction "Ref.SiAct"
 "DopingConcentration" "MaxTransDiff" 1)
(sdedr:definerefinementplacement
"RefPlace.SiAct"
 "Ref.SiAct" "RWin.Act" )
; Active above STI
(sdedr:definerefinementwindow "RWin.ActSTI"
 "Rectangle"
 (position     ( Xmax 1.6)  0.3    0.0)
 (position      Xmax        0.4    0.0 ))
 (sdedr:definerefinementsize "Ref.SiActSTI"
  0.05         0.01
  0.01          Gpn )
(sdedr:definerefinementfunction "Ref.SiActSTI"
 "DopingConcentration" "MaxTransDiff" 1)
(sdedr:definerefinementplacement
"RefPlace.SiActSTI"
 "Ref.SiActSTI" "RWin.ActSTI" )
; Retro start
(sdedr:definerefinementwindow "RWin.Retro"
 "Rectangle"
 (position     0.0      0.4    0.0)
 (position     Xmax      0.6     0.0 ))
 (sdedr:definerefinementsize "Ref.Retro"
  0.1          0.015
  0.05          0.0075 )
(sdedr:definerefinementfunction "Ref.Retro"
 "DopingConcentration" "MaxTransDiff" 1)
(sdedr:definerefinementplacement "RefPlace.Retro"
 "Ref.Retro" "RWin.Retro" )

; Po Gate Multibox
(sdedr:definerefinementwindow
"MBWindow.Gate"
 "Rectangle"
 (position 0.0 Ypol  0.0)
 (position Xg  Ygox  0.0) )
(sdedr:definemultiboxsize "MBSize.Gate"
  0.015  (/ Hpol 4.0)
  0.01      2e4
  1.0     1.35 )
```

```
(sdedr:definemultiboxplacement "MBPlace.Gate"
 "MBSize.Gate"  "MBWindow.Gate" )
; GateOx Multibox
(sdedr:definerefinementwindow
"MBWindow.GateOx"
 "Rectangle"
 (position     0.0       Ygox  0.0)
 (position (+ Xg  0.025)  0.0   0.0) )
(sdedr:definemultiboxsize "MBSize.GateOx"
  0.015     0.0005
  0.008      1e4
  1.0      1.35 )
(sdedr:definemultiboxplacement
"MBPlace.GateOx"
 "MBSize.GateOx"  "MBWindow.GateOx" )
; GateOx warum zwei mal ????
(sdedr:definerefinementsize "Ref.GOX"
  0.015     0.0005
  0.003     0.0002 )
(sdedr:definerefinementregion "RefPlace.GOX"
 "Ref.GOX" "R.Gateox" )
; Channel Multibox
(sdedr:definerefinementwindow
"MBWindow.Channel"
 "Rectangle"
 (position        0.0         0.0   0.0)
 (position   (+ Xg  0.025)    0.05  0.0) )
(sdedr:definemultiboxsize "MBSize.Channel"
  0.005     0.01
  0.002     1e4
  1.0       1.35 )
(sdedr:definemultiboxplacement
"MBPlace.Channel"
 "MBSize.Channel" "MBWindow.Channel" )

; Save CMD file
(sdedr:writecmdfile "n475_half_msh.cmd")
;(sdedr:writecmdfile "n475_msh.cmd")

; Build Mesh
(system:command "mesh F tdr n475_half_msh")
;(system:command "mesh F tdr n475_msh")

; Reflect device

(system:command "tdx mtt x ren drain=source
n475_half_msh n475_msh")
```

(B2) Dynamic Simulation Command File --

```
#setdep @node|-1:all@

#- Sentaurus Device input deck for a transient
mixed-mode simulation of the
#- switching of an inverter build with a nMOSFET
and a pMOSFET.

# define the n-channel MOSFET;
Device NMOS {

  Electrode {
    { name="source"    Voltage=0.0 Resist=20
Area=2.7 }
    { name="drain"     Voltage=0.0 Resist=20
Area=2.7 }
    { name="gate"      Voltage=0.0 Resist=20
Area=2.7 }
    { name="substrate" Voltage=0.0
Resist=@Rwell@ Area=2.7 }
    { Name="substrate_mirrored" Voltage=0.0
Resist=@Rwell@ Area=2.7}
    { name="bottom"    Voltage=0.0 Area=2.7 }
  }

File {
        grid    = "@tdr@"
              Plot    = "nmos"
              Current = "nmos"
        Parameter="@parameter@"
}

  Physics {
   Hydrodynamic( eTemperature )
   EffectiveIntrinsicDensity( oldSlotboom )
  }

  Physics(Material="Silicon"){
   eQuantumPotential

   Mobility(
     PhuMob
     eHighFieldSaturation( CarrierTempDrive )
     hHighFieldSaturation( GradQuasiFermi )
     Enormal
   )
   Recombination(
     SRH( DopingDep )
     Band2Band
     eAvalanche(CarrierTempDrive)
     hAvalanche(GradQuasiFermi)
   )
  }
}

# define the p-channel MOSFET;
Device PMOS{
  Electrode {
    { Name="source"    Voltage=0.0 Resist=20
Area=3.1 }
    { Name="drain"     Voltage=0.0 Resist=20
Area=3.1 }
    { Name="gate"      Voltage=0.0 Resist=20
Area=3.1 }
    { Name="substrate" Voltage=0.0
Resist=@Rwell@ Area=3.1 }
    { Name="substrate_mirrored" Voltage=0.0
Resist=@Rwell@ Area=3.1}
    { name="bottom"    Voltage=0.0 Area=2.7 }
  }

  File {
        grid    = "@tdr:+1@"
              Plot    = "pmos"
              Current = "pmos"
        Parameter="@parameter@"
  }

  Physics {
   Hydrodynamic( hTemperature )
   EffectiveIntrinsicDensity( oldSlotboom )
  }

  Physics(Material="Silicon"){
   hQuantumPotential

   Mobility(
     PhuMob
     hHighFieldSaturation( CarrierTempDrive )
     eHighFieldSaturation( GradQuasiFermi )
     Enormal
   )
   Recombination(
     SRH( DopingDep )
     Band2Band
     hAvalanche(CarrierTempDrive)
     eAvalanche(GradQuasiFermi)
   )
  }
}

# definition of input signal time[s]
System {
  Vsource_pset v0 (n1 n0) { pwl = (0.0e+00 0.0
                               10e-12 0.0
                               45e-12 @Vdd@
                             1200e-12 @Vdd@
                             1222e-12 0.0
                             2400e-12 0.0
                             2435e-12 @Vdd@
                             3600e-12 @Vdd@
                             3622e-12 0.0
                             4800e-12 0.0
                             4835e-12 @Vdd@
                             5000e-12 @Vdd@)}

# definition of SPICE enviroment (nodes)

  NMOS nmos( "source"=n0 "drain"=n3 "gate"=n1
"substrate"=n0 "substrate_mirrored"=n0
"bottom"=n4)
  PMOS pmos( "source"=n2 "drain"=n3 "gate"=n1
"substrate"=n2 "substrate_mirrored"=n2
"bottom"=n5)
  Capacitor_pset c1 ( n3 n0 ){ capacitance = 1.35e-
14 }

  Set (n0 = 0)
  Set (n2 = 0.0)
  Set (n3 = 0.0)
  Set (n4 = 0.0)
  Set (n5 = 0.0)

  Plot "nodes.plt" (time() n0 n1 n2 n3 n4 n5)
}

File {
   Current= "inv"
   Output = "inv"
}

Plot {
     eDensity hDensity
   eCurrent hCurrent
   ElectricField eEnormal hEnormal
   eQuasiFermi hQuasiFermi
   Potential Doping SpaceCharge
   DonorConcentration AcceptorConcentration
   AvalancheGeneration
}

CurrentPlot { Potential ((0 0.3)) }

Math {
  Extrapolate
  Avalderivatives
  Derivatives
  RelErrControl
  Digits=5
  *Notdamped=50
  *Iterations=12
  NoCheckTransientError
}

#-build up initial solution

Solve {
  NewCurrentPrefix = "ignore_"
   Coupled(Iterations=100
LineSearchDamping=1e-4){ Poisson
nmos.eQuantumPotential
```

```
        pmos.hQuantumPotential}
  Coupled { Poisson nmos.eQuantumPotential
pmos.hQuantumPotential Electron Hole }
  Coupled { Poisson nmos.eQuantumPotential
pmos.hQuantumPotential
  Electron nmos.eTemperature
  Hole pmos.hTemperature}
  Coupled { Poisson nmos.eQuantumPotential
pmos.hQuantumPotential Electron
  nmos.eTemperature Hole pmos.hTemperature
Contact Circuit}

  Unset (n2)
  Unset (n3)
  Unset (n4)
  Unset (n5)

  Quasistationary ( InitialStep=2.77777e-4
MaxStep=0.02 MinStep=0.00001
                Goal { Node=n2 Value=@Vdd@ }
                Goal { Node=n3 Value=@Vdd@ }
                Goal { Node=n4 Value=@Vbn@ }
                Goal { Node=n5
Value=@Vbp@ }                    )

     {Coupled (iterations=10) { Poisson
nmos.eQuantumPotential pmos.hQuantumPotential
Electron
 nmos.eTemperature Hole pmos.hTemperature
Contact Circuit} }
```

```
   NewCurrentPrefix = ""

  Set (n2 = @Vdd@)
  Set (n3 = @Vdd@)

  Unset (n3)
  Unset (n4)
  Unset (n5)

# finall dynamic simulation

   Transient (
    InitialTime=0 FinalTime=5000e-12
    InitialStep=1e-12 MaxStep=1e-9 MinStep=0.5e-
15
    Increment=1.3
 )
   { Coupled (iterations=20){ nmos.poisson
nmos.eQuantumPotential nmos.electron  nmos.hole
   nmos.eTemperature nmos.contact
   pmos.poisson pmos.hQuantumPotential
pmos.hole  pmos.electron pmos.hTemperature
pmos.contact
           circuit }
     Plot (time=(0; 100e-12) NoOverwrite)
   }
}
```

(B3) Silicon Parameter File --

```
* Generation & Recombination:
* Recombination( SRH( DopingDep Tunneling ) )

Scharfetter * relation and trap level for SRH
recombination:
{ * tau = taumin + ( taumax - taumin ) / ( 1 +
( N/Nref )^gamma)
  * tau(T) = tau * ( (T/300)^Talpha )
(TempDep)
  * tau(T) = tau * exp( Tcoeff * ((T/300)-1) )
(ExpTempDep)
          taumin   = 0.0000e+00 ,    0.0000e+00
          # [s]
*         taumax   = 1.0000e-05 ,    3.0000e-06
          # [s]
*         taumax   = 1.0000e-06 ,    1.0000e-06
          # [s]
          taumax   =    2.1e-6 ,          7e-7      # [s]
*         Nref      = 1.0000e+16 ,    1.0000e+16
          # [cm^(-3)]
          Nref      = 1.0000e+17 ,    1.0000e+17
          # [cm^(-3)]
```

```
          gamma   = 1 ,       1          # [1]
          Talpha   = -1.5000e+00 ,  -1.5000e+00
          # [1]
          Tcoeff   = 2.55 ,  2.55       # [1]
          Etrap     = 0.0000e+00      # [eV]
}

TrapAssistedTunneling * lifetimes:
{ * See Dessis manual `Trap-Assisted
Tunneling/SRH'
          S           = 3.5      # [1]
          hbarOmega          = 0.068  # [eV]
          MinField             = 1.0000e+03      #
[V/cm]
          m_theta = 0.258 ,           0.24       # [1]
          Z           = 0.0000e+00      # [1]
}

Recombination( Band2Band )

Band2BandTunneling
{ * See Dessis manual `Band-To-Band Tunneling'
```

```
  * min potential difference on length dPot/E (for
traditional & Hurkx models):
            dPot      = 1.1      # [V]
}
* Mobility Models

* Lucent Mobility Model
* Parameters according to
* M.N. Darwish et al., IEEE Trans. Elect. Dev. 44,
No 9., p 1529-1537, 1997

EnormalDependence:
{ * mu_Enorm^(-1) = mu_ac^(-1) + mu_sr^(-1)
with:
  * mu_ac = B / Enorm + C (T/T0)^(-k)
(N/N0)^lambda / Enorm^(1/3)
  * mu_sr^-1 = Enorm^(A+alpha*n/(N+N1)^nu) /
delta + Enorm^3 / eta
  * EnormalDependence is added with factor exp(-
l/l_crit), where l is
  * the distance to the nearest point of
semiconductor/insulator interface.
  * Factor is equal to 1 if l_crit > 100.
            B         = 3.61e+07 ,        1.51e+07
            # [cm/s]
            C         = 1.70e+04 ,        4.18e+03
            # [cm^5/3/(sV^2/3)]


            N0        = 1           ,        1
            # [cm^-3]
            lambda  = 0.077   ,         0.0158
            # [1]
            k         = 1.7        ,        0.9
            # [1]
            delta     = 3.58e+18 ,        4.10e+15
            # [V/s]
            A         = 2.582    ,        2.118     # [1]
*           alpha     = 6.85e-21 ,        7.82e-21# [1]
* Increase to accomodated for Density Gradient
Model
            alpha     = 2e-20    ,        3e-20     # [1]
            N1        = 1           ,        1
            # [cm^-3]
            nu        = 0.0767   ,        0.123
            # [1]
            eta       = 5.82e+30 ,        2.0546e+30
            # [V^2/cm*s]
            l_crit    = 1        ,        1
            # [cm]
}

* Highfield Saturation

* Transport Models
* Hydrodynamics
EnergyRelaxationTime
```

```
{ *  Energy relaxation times in picoseconds
            (tau_w)_ele        = 0.3   # [ps]
            (tau_w)_hol        = 0.3     # [ps]
}

HighFieldDependence:
{ * Caughey-Thomas model
  * and HydroHighField mobility is used.
            K_dT    = 1e-4,   1e-4 # [1]
  *         beta0   = 2.02 , 2.25
            beta0   = 2.02 , 2.25

  *         vsat0     = 0.897e+07 ,      8.3700e+06
            # [1]
            vsat0     = 0.91e+7   ,        8.13e6   # [1]
}

AvalancheFactors
{ *  Coefficientss for avalanche generation with
hydro
  *  Factors n_l_f, p_l_f for energy relaxation length
in the expressions
  *  for effective electric field for avalanche
generation
  *  eEeff = eEeff / n_l_f  ( or b = b*n_l_f )
  *  hEeff = hEeff / p_l_f  ( or b = b*p_l_f )
  *  Additional coefficients n_gamma, p_gamma,
n_delta, p_delta
            n_l_f     = 0.87    # [1]
            p_l_f     = 0.87    # [1]
            n_gamma          = 1         # [1]
            p_gamma          = 1         # [1]
            n_delta  = 1.5      # [1]
            p_delta  = 1.5      # [1]
}
EnergyFlux
{ *  Coefficient in front of the energy flux equation
  *  energy_flux_coef=0.6 corresponds to Stratton
model
            *energy_flux_coef_ele       = 0.6      # [1]
            *energy_flux_coef_hol       = 0.6      # [1]
  *  energy_flux_coef=1.0 corresponds to Blotekjear
model
        energy_flux_coef_ele            = 1.0      # [1]
            energy_flux_coef_hol        = 1.0      # [1]
}

ThermalDiffusion
{ *  Thermal diffusion factor, td:
td*mu*kB*N*grad(T), td=0 for Stratton model

            *td_n     = 0.0000e+00      # [1]
            *td_p     = 0.0000e+00      # [1]

  *  Thermal diffusion factor, td:
td*mu*kB*N*grad(T), td=1 for Blotekjear model
```

```
        td_n          = 1.0000e+00      # [1]
              td_p          = 1.0000e+00      # [1]

*  Carrier diffusion factor, td_g:
mu*kB*(td_g*Tcarrier + (1-
td_g)*Tlattice)*grad(N)
              td_gn    = 1         # [1]
              td_gp    = 1         # [1]
}
* Density Gradient Quantum Transport
```

```
QuantumPotentialParameters
{ * gamma:  weighting factor for quantum potential
  * theta:  weight for quadratic term
  * xi:     weight for quasi Fermi potential
  * eta:    weight for electrostatic potential
            gamma   = 3.6 ,   5.6        # [1]
            theta   = 0.5 ,   0.5        # [1]
            xi      = 1 ,     1          # [1]
            eta     = 1 ,     1          # [1]
}
```