EPFL

Thèse n° 10 483

# Adaptive Body Biasing in Strong Body Factor Technologies

Présentée le 29 janvier 2021

Faculté des sciences et techniques de l'ingénieur
Laboratoire de circuits pour télécommunications
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

**Thomas Christoph MÜLLER**

Acceptée sur proposition du jury

Prof. S. Carrara, président du jury
Prof. A. P. Burg, Dr M. Pons Solé, directeurs de thèse
Prof. J. Rodrigues, rapporteur
Prof. L. Koskinen, rapporteur
Prof. M. Shoaran, rapporteuse

École polytechnique fédérale de Lausanne

2021

# Acknowledgements

I am very grateful to my thesis supervisor Prof. Andreas Burg for giving me the opportunity to do my PhD at the Telecommunication Circuits Laboratory at EPFL. The countless technical discussions pushed me forward and his valuable comments sharpened this work which would not exist in this form otherwise. Thanks for the continuous unconditional support and guidance as well as for providing such an incredible interdisciplinary lab environment. Further, I would like to express my gratitude to my thesis co-director at CSEM, Dr. Marc Pons Solé: thanks for the dedication in supporting me throughout these years and for all the invaluable advice.

I would like to thank the committee of my thesis defense: Prof. S. Carrara of the Integrated Circuits Laboratory at EPFL who acted as the president of the jury and Prof. Mahsa Shoran from the Integrated Neurotechnologies Laboratory of EPFL who acted as the internal expert. In addition I would like to to thank the two external examiners: Prof. Joachim Rodrigues from the Integrated Electronic Systems group of Lund University and Prof. Lauri Koskinen from the Department of Future Technologies of Turku University.

Further, I would like to express my gratefulness to all my colleagues at CSEM I worked closely with: my supervisor Stéphane Emery, as well as Daniel Séverac, David Ruffieux, Erfan Azarkhish, Ernesto Pérez Serna, Jean-Luc Nagel, Loïc Zahnd, Regis Cattenoz, Stéphane Devise, Themis Mavrogordatos, and Virginie Moser - thanks for all the support and technical discussions. I also would like to thank USJC for providing the funding, silicon, and early technology access for the CSEM project which lead to this thesis.

I wish to show my gratitude to all former and current members of the Telecommunication Circuit Laboritory: Adam Teman, Adrian Schumacher, Alexios Balatsoukas-Stimming, Andrea Bonetti, Andreas Kristensen, Andrew Austin, Christian Senning, Jeremy Constantin, Lorenz Schmid, Matthieu Cotting, Nicholas Preyss, Pascal Giard, Orion Afisiadis, Reza Ghanaatian Jahromi, Robert Giterman, Ricardo Gomez Gomez, Sitian Li, and Shrikanth Ganapathy - without you this time would not have been half as enjoyable and I would not have learned half as much without your incredible talent and knowledge in such diverse fields. A special thanks goes to our secretary Ioanna Paniara for keeping our back free from all the organisational work.

Finally, my family supported and believed in me over all these years of education: I wholehartedly would like to thank my mother Heike, my father Mathias as well as my brothers Carsten, and Constantin.

*Lausanne, December 18, 2020* C. M.

# Abstract

With the advent of intelligent sensor nodes in everyday life, low power aspects of system design become more and more important. Adaptive body biasing is a promising methodology to achieve dynamic adaptation of the tradeoff between performance and energy by shifting the threshold voltage of transistors in a digital design. This approach, in combination with a low supply voltage, provides a strong knob to the designer to rapidly shift the circuits operating point from deep sub threshold operation for slow and low leakage retention, to fast, higher performance operation to for short, but demanding tasks. This thesis concentrates on such designs using the deeply depleted channel technology in which body control is particularly effective.

The first part of this thesis is dedicated to strategies and tools supporting the digital design process of circuits using adaptive body biasing. A methodology to compare a standard cell library characterised in different operating points, defined by supply voltage, process corner, temperature, and bias points is presented first.

Next, we present a methodology to exhaustively and rapidly map out the supply-voltage/bias-voltage design space using a heavily pruned cell library. We extract speed and power of a simple example design across the entire design space and show a methodology to scale the characteristics of the small reference up to a more complex design. In a case study this modelling approach achieves an error of less than 1% on the total power relative to an actual characterisation of the full library at the same design point.

The second part of the thesis analyses three chips implementing different biasing schemes in USJC 55nm DDC. The first two were designed by CSEM with components and measurements contributed from this thesis while the third one was entirely designed for this thesis.

The first chip utilises a biasing scheme based on the first order approximation that the circuit speed is proportional to the on-current which can be driven by PMOS and NMOS transistors. This is implemented using an analog control loop, setting both PMOS and NMOS on-currents equal to a reference current provided by current DAC. The SoC characterisation is presented with the objective of identifying suitable operating modes and bias points, including a reliability and retention analysis of the SRAM. A series of ring oscillators constructed from the most common standard cells has also been integrated and provides measurement support for the first part of the thesis.

The second chip extends the Calanda biasing scheme with a secondary regulation loop that is based on an FLL, designed in this thesis, in combination with a configurable standard cell based ring oscillator. The user can directly program a target frequency and the biasing system

regulates the current DAC accordingly. We show that this approach effectively overcomes the drawbacks of the current based approach resulting in an effective regulation.
Finally the third chip presents a novel biasing scheme that was designed in this thesis and is tailored toward simplicity. It utilises two constant voltages for the PMOS bias to switch between retention and operation. A charge pump controlled by a standard cell compatible distributed on current balance sensor regulates the NMOS bias such that the on-currents of PMOS and NMOS match. We show that this simple approach, in conjunction with a well chosen operating point, can be efficient across corners.

**Key Words:** Digital VLSI Circuits, Low-Power Design,
Deeply Depleted Channel, Adaptive Body Biasing, PVT Compensation,
Standard Cell Library Analysis, Design Space Exploration, Throughput Scaling, Retention, Bias Generation.

# Zusammenfassung

Mit der Verbreitung von intelligenten Sensor-Knoten im Alltag und den Fortschritten im Feld des maschinellen Lernens wird energiesparendes Systemdesign zunehmend bedeutender. Adaptives Body Biasing ist dabei eine vielversprechende Methode um eine dynamische Anpassung zu erreichen indem die Schwellenspannung der Transistoren im digitalen Design verschoben wird. Dieser Ansatz, zusammen mit einer niedrigen Versorgungsspannung, gibt dem Designer ein mächtiges Werkzeug um den Betriebspunkt schnell von weit unter der Schwellspannung mit niedrigen Leckströmen zu einem Punkt mit höherer Leistung für kurze, aber anspruchsvolle Aufgaben zu verschieben. Diese Arbeit konzentriert sich auf Schaltungen in USJC 55nm DDC wo Body Biasing besonders effektiv ist.

Der erste Teil dieser Arbeit beschäftigt sich mit dem Entwicklungsprozess für digitale Schaltungen unter der Anwendung von Adaptivem Body Biasing. Es wird eine Methode entwickelt um eine Bibliothek von Standardzellen welche unter verschiedenen Betriebspunkten (Temperatur, Versorgungsspannung, Prozess-Corner) charakterisiert wurde gegeneinander zu vergleichen.

Als nächstes präsentieren wir eine Methode mit der der Raum der durch die Variation der Body- und Versorgungsspannung aufgespannt wird unter Verwendung einer stark beschnittenen Zellbibliothek kartiert werden kann. Wir extrahieren mit einem einfachen Refererenzdesign die dynamischen und statischen Ströme über den gesamten Designraum und zeigen eine Methodik zur Skalierung auf ein komplexeres Design. In einer Fallstudie wird einen Fehler von weniger als 1% des Gesamtenergieverbrauchs erreicht.

Der zweite Teil dieser Arbeit beschäftigt sich mit drei Chips die drei verschiedene Biasverfahren implementieren. Die ersten beiden wurden von einem Team innerhalb von CSEM entworfen wobei Teilkomponenten aus dieser Arbeit beigetragen wurden. Der dritte Chip wurde speziell für diese Arbeit entworfen.

Calanda verwendet ein Bias-Verfahren auf Basis der Annäherung das die Schaltgeschwindigkeit proportional zum Sättigungsstrom der PMOS- und NMOS-Transistoren ist. Eine analoge Regelschleife reguliert dabei über den Biaskontakt die PMOS- und NMOS-Ströme auf äquivalent zu einem Referenzstrom der von einem Digital-Analogwandler mit Stromausgang gesetzt wird. SRAM und Core werden mit dem Ziel charakterisiert geeignete Betriebspunkte über einen weiten Betriebsbereich zu finden. Zusätzlich wurde eine Schar von Ringoszillatoren integriert um die Charakterisierung und Modellierung aus dem ersten Teil dieser Arbeit mit Messdaten zu unterfüttern.

Nakayama erweitert das Bias-Verfahren von Calande um eine sekundäre Regelschleife auf Basis einer FLL, welche als Teil dieser Arbeit entworfen wurde, zusammen mit einem Ringoszillator

auf Standardzellbasis. Die Regelschleife kontrolliert die DAC-Codes so das eine vom Anwender gesetzte Zielfrequenz erreicht wird. Wir zeigen das dieser Ansatz in der Lage ist die Nachteile des Calanda-Konzeptes zu umgehen und eine effektive Regulierung implementiert.
Snaefellsjokull nutzt ein Biaskonzept welches auf eine besonders simple Implementierung setzt wobei zwei konstante Biasspannungen für die PMOS-Transistoren zwischen einen Haltemodus und Operationsmodus definieren. Eine Ladungspumpe wird von einem Standardzellkompatiblen Strombalance-Sensor kontrolliert um die NMOS-Biasspannung so zu regulieren das der Sättigungsstrom von PMOS und NMOS identisch sind. Wir zeigen das dieser einfache Ansatz zu einer effizienten Regulierung führen kann.

**Schlüsselworte:** Digitale Schaltungen, Low-Power Design,
Deeply Depleted Channel, Adaptives Body Biasing, PVT Kompensation,
Standardzellbibliothek-Analyse, Design Space Exploration, Durchsatz-Skalierung, Stromspar-Modus, Bias-Generator.

# Contents

# 1 Introduction

Over the last years the requirements of wireless sensors and actors have shifted significantly. What was once a simple "dumb" sensor is now expected to integrate into smart systems and environments, allowing for an orchestrated control. This development is generally subsumed under the buzzword *Internet of Things*, a term initially coined by Kevin Ashton in the context of using RFID for supply chain tracking [1]. The meaning has since broadened to include all kinds of autonomous devices which are by some method connected to a network. A typical example would be an intelligent heating system using wireless temperature and gas sensors across the apartment together with window sensors to make "clever" decisions on whether to turn on the heating and ventilation or not. Typically these devices are of ubiquitous nature and are operated with a fairly low consciousness on our side: We expect these devices to operate reliably, but stay out of our way when not needed with easy installation, typically meaning that operation is battery powered.

Hence, the key requirement for these sensors is a very low power footprint and autonomous operation over a long time: The direct interaction with these kind of devices is so rare that, contrary to highly interactive devices like smart phones, a frequent recharge or battery replacement is not acceptable.

Typically these kind of devices integrate some kind of sensor, together with some ability to preprocess data locally and some wireless protocol for communication. Figure 1.1 illustrates the power profile of these components: Sampling data from the sensor is typically the lowest power contributor while the most power expensive part is active communication [2]. This leaves a nice tradeoff in between where local processing, often applying the pattern matching capabilities from artificial intelligence algorithms, are used whenever possible to reduce the number of necessary transmissions.

A prime example for this kind of architecture is the implementation of voice control: a low power wakeup circuit is capable to detect a keyword (Alexa, Ok Google, Hey Siri,... ) which then triggers a transmission of the speech command into the cloud only when necessary.

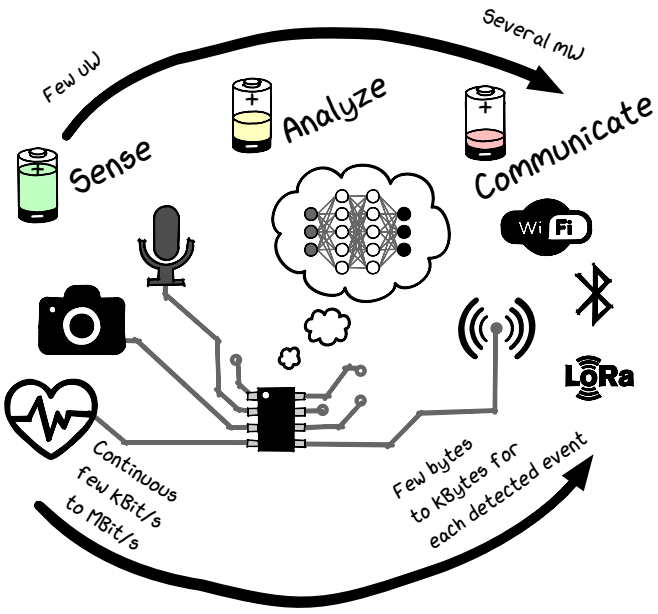



Figure 1.1 – Cost of interaction: Sensing vs. Analysis vs. Data transmission

## 1.1 Low Power Design

Achieving the associated ultra low power requirements sets a particular challenge for the SoC designer: the circuit must be able to idle at as low as possible power for the majority of time, but must also be able to achieve the performance necessary to apply the, often demanding [3], filtering or pattern detection algorithms required for its particular task. The answer typically is to add some kind of adaptivity to the circuit, shifting the mode of operation back and forth between a high performance computation mode and an ultra low power idle mode [4]. Circuit power can be broken down into two main components:

**Dynamic Power** is the power due to the switching activity in the chip, comprising the power attributed to the charge and discharge processes of wire and gate capacitances. A first order approximation can be derived as follows: $V_{DD}$ is the supply voltage and $f$ the operating frequency. $C_L$ represents the total load capacitance over all switched transistors, which consists mainly of the gate capacitance as well as wire routing capacitances. The switching factor $\alpha_{sw}$ represents an activity factor, considering that a) not all transistors are switched all the time and b) that energy spent to charge a gate capacitance during a low to high transition will be dissipated again in the next high to low transition.

$$P_{dyn} \quad = \quad \alpha_{sw} \cdot C_L \cdot V_{DD}^2 \cdot f \tag{1.1}$$

**Leakage Power**, also known as static power, subsumes all components resulting from the

currents $I_{OFF}$ that flow even when there is no activity inside the circuit.

$$P_{leak} = V_{DD} \cdot I_{OFF} \tag{1.2}$$

The main contributor for the $I_{OFF}$ are residual sub-threshold currents through "closed" MOSFETs, followed by substrate leakages through reversed p-n-junctions used in the construction of the devices [5].

### 1.1.1 Reducing Dynamic Power

In principle all parameters in (1.1) and (1.2) can be optimisation targets to reduce circuit power. For example, the activity $\alpha_{sw}$ can be tackled with all kinds of dynamic power management techniques, with the most prominent being clock gating [6]–[8], a design time methodology where logic is introduced into the clock tree in a carefully designed manner to only clock the parts of the circuit that are currently in use. Similarly, the clock frequency $f$ can be adjusted on those to reduce the switching in a given time period. Adjusting the supply voltage $V_{DD}$ is very promising due to the squared contribution towards $P_{dyn}$.

However, the circuit speed is defined by the rate at which the gate capacitances can be charged and discharged which can be modelled in first order approximation by a simple RC model. Taking the simple approximation of the charge- and discharge speed based on the on-resistance [9, p. 200ff.] and, for the sake of simplicity, even further simplifying it by removing the channel length modulation term and assuming the PMOS and NMOS currents are well balanced we can show that the propagation delay is in first order approximation are proportional to the input voltage $V_{DD}$ and the saturation current $I_{DSAT}$. For the purpose of this work we will assume the latter to be closely approximated by the on-current $I_{ON}$ which we define as the drain source current $I_{DS}$ imposed by the MOSFET for the case of $V_{GS} = |V_{DS}| = V_{DD}$, i.e. the current flowing when we just started the (dis)charge of the output node while also assuming that the gate voltage has completely settled. With this we can formulate the basic model for the propagation delay

$$tp \propto C_L \cdot V_{DD} \cdot \frac{1}{I_{ON}}, \tag{1.3}$$

or, when considering the achievable frequency:

$$f \propto \frac{I_{ON}}{C_L \cdot V_{DD}}. \tag{1.4}$$

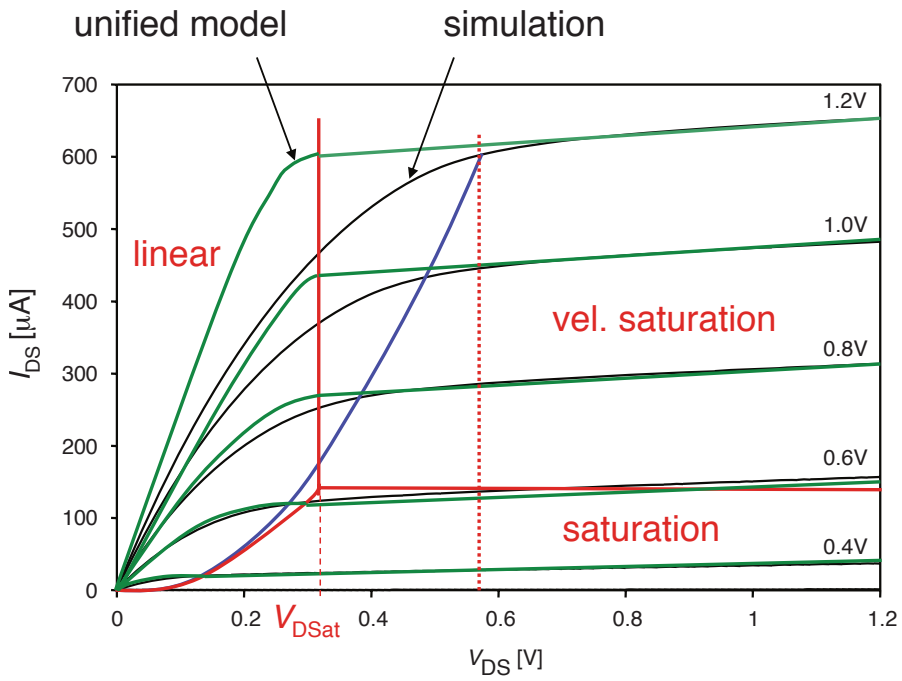



Figure 1.2 – Estimated current $I_{DS}$ (green) vs. actual current (black) from the spice BSIM-4 model for different gate source voltages. Reproduced from [10, p. 28].

$I_{DS}$ is indeed is a function of the Supply voltage as shown in the unified empirical model described in Rabaey et al. [9, p. 101]:

$$I_{DS} = 0 \text{ for } V_{GT} \leq 0 \tag{1.5}$$

$$I_{DS} = k' \frac{W}{L} \left( V_{GT} V_{min} - \frac{V_{min}^2}{2} \right) (1 + \lambda V_{DS}) \text{ for } V_{GT} \geq 0 \tag{1.6}$$

$$\text{with } V_{min} = min\left(V_{GT}, V_{DS}, V_{DSAT}\right), \tag{1.7}$$

$$\text{and } V_{GT} = V_{GS} - V_T \tag{1.8}$$

The supply voltage appears in this equation mainly as the gate source voltage $V_{GS}$ (as the input of a digital cell) and as the drain source voltage $V_{DS}$ for driving the load current. The remaining parameters are the width $W$ and length $L$ of the transistor, the process transconductance $k'$, and the channel length modulation parameter $\lambda$. Finally the saturation voltage $V_{DSAT}$ denotes the point at which the velocity of the current driving carriers no longer linearly increases with the strength of the field imposed by $V_{GS}$, but saturates.

This model is simplified, but, as shown in Fig. 1.2, predicts the current $I_{DS}$ for a large $V_{DS}$ and $V_{GS}$ fairly well, which is the region the transistor in a digital CMOS gate operate for a large share of the charge/discharge time [10, p. 28].

For short channel transistors, which are typically used in digital gates due to space constraints,

the current is typically limited by velocity saturation, resulting in a linear relation between $V_{GS}$ and $I_{DS}$. We can observe that in Fig. 1.2, where starting from 600 mV upwards every gate voltage increase of 200 mV results in an increase in current of approximately 170 µA. For supply voltages below 600 mV we reach the saturation region where the relationship between $V_{GS}$ and $I_{DS}$ becomes quadratic.

While (1.3) on first glance suggests a slowdown with an increased $V_{DD}$ we observe an effect of $V_{DD}$ on $I_{DS}$ which is significantly larger: considering Fig. 1.2 a mere 200 mV step in $V_{GS}$ from 600 mV to 800 mV results in an increase of current from roughly 100 µA to about 270 µA. Hence, we can use the drive current $I_{DS}$ of the MOSFET as a simplified proxy for the circuit speed. As such we can expect a linear dependence between the achievable operating frequency of a circuit and the supply voltage in velocity saturation and a quadratic dependence at lower supply voltage when the transistors are operated mostly in saturation during the charge and discharge of the capacitive load of the gate.

Consequently it is necessary to regulate the supply voltage and clock frequency carefully and jointly to keep a balance between not introducing timing violations and, in the opposite direction, not commencing excessive margins. This adaptive voltage and frequency scaling (ADVFS) approach is very today widely adopted in the industry. As this approach effects both the squared $V_{DD}$ and the frequency term in (1.1) the designer is left with a very powerful tool for significant power reduction in cases where the workload is lower than the achievable throughput. Not surprisingly this approach is commonly used in high performance CPUs to reduce idle power [11], but has also been used in low power SoC proposals such as [12].

However, once the supply voltage dips below the threshold voltage the current driving capabilities reduce drastically, converging towards, but never reaching the approximation of zero from the unified model in (1.5). Instead, the following model for weak inversion conduction can be used which shows an exponential dependence of the drive current $I_D$ of the gate source voltage $V_{GS}$ and the drain source voltage $V_{DS}$ [9, p. 103]:

$$I_{DS} = I_S \cdot e^{\frac{V_{GS}}{n \cdot kT/q}} \left(1 - e^{-\frac{V_{DS}}{kT/q}}\right) \text{ for } V_{GT} \leq 0 \tag{1.9}$$

Pushing a circuit into this sub-threshold domain can be very effective: Myers et al. [13] show a reduction in active power of more than 5000x when sweeping the supply voltage from 1.2 V to 250 mV at which point it reaches the same order of magnitude as the circuit leakage. The exponential reduction in current driving capability obviously has a massive effect on the propagation delay: the circuit speed reduces from 66 MHz down to 27 kHz, a factor of 2444x. Furthermore, doping variability induced shifts of the threshold voltage between different devices on the same chip result in a significant shift of drive capability on the exponential slope.

### 1.1.2 Reducing Leakage Power

While the sub-threshold current variation has a significant impact on the speed of circuits operating with a supply voltage below the threshold voltage $V_T$, it is also an important factor for circuits operating with above the threshold. In this case, the leakage current consumed when transistors are not switching is determined by (1.9).

This means that just a small amount of shift in either the device threshold voltage or the gate source voltage can result in a substantial shift in leakage current due to the exponential nature of this curve. The result is, that a few particularly bad devices can dominate the overall circuit leakage [5, p. 9].

### 1.1.3 Power versus Energy

Maximizing the battery life of a design is equal to minimzing the energy consumed by the design. Hence, an operation close to the minimum energy point (MEP)[14] is desirable to maximize the operation time. Power is the rate of energy consumed per time and is typically used for measurements as it can be easily derived from the supply currents and voltage. When the operating frequency of a device varies, power and energy are not the same and the energy minimum point may be significantly different from the minimum power. However, for constant frequency applications with no idle periods (as considered in this thesis), minimizing energy corresponds to minimizing power. Hence, power and energy are used interchangeably in this thesis.

## 1.2 Variation Mitigation Strategies for Low Power Designs

When aggressively scaling the supply voltage down into the near- and sub-threshold domain the device becomes particularly susceptible for random process, supply voltage, and temperature variation. In [15] Dreslinski et al. show how the gate delay variation due to process degrades from 1.3X at nominal voltage over 5X at near-threshold up to 14X in the sub-threshold domain. While the former can be easily covered by adding margins using a multi mode multi corner (MMMC) approach for the static timing analysis (STA) at design time, the latter two cases introduce highly pessimistic margins that degrade the circuit characteristics for corners other than the worst case.

Those margins can be avoided by adding a control loop that adjusts either the supply voltage or the circuit clock frequency at run-time so that it matches the specific (rather than the worst case) operating conditions. More advanced schemes include even rapid adaptation to the supply voltage ripple of a simple DCDC [16] or even adaptation of the clock frequency to data dependent circuit delay variations [17].

An alternative control approach is the use of body biasing [18]. While aggressive voltage scaling

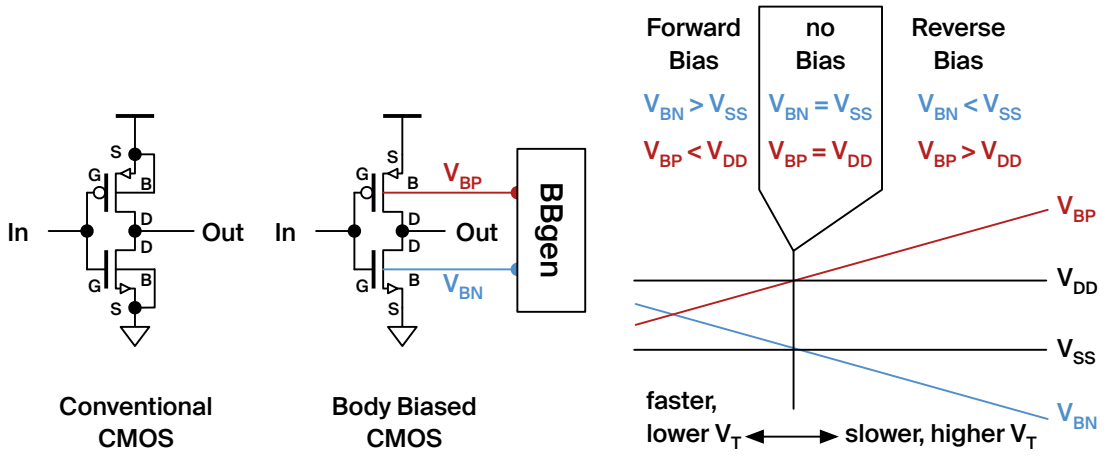



Figure 1.3 – Left: Inverter with conventional CMOS in comparison to an inverter with body biasing. A body bias generator (BBgen) produces the two body voltages $V_{BP}$ and $V_{BN}$, setting the source bulk voltage $V_{SB}$ of NMOS and PMOS respectively. Right: forward and reverse biasing ranges for NMOS and PMOS.

shifts the circuit supply voltage relative to the threshold voltage, body biasing achieves a similar effect by shifting the threshold voltage relative to the supply by applying a bias voltage to the substrate, resulting in a modulation of the tradeoff between leakage and circuit speed.

The above described dynamic adaptation requires a sensor to identify the circuit speed under the given operating conditions to adjust the available control knobs accordingly. Several sensing approaches have been proposed in literature: The first category tries to estimate the general speed of the die, for example by counting cycles of ring oscillators [19] or measuring the propagation through critical path replicas [20], [21] or canary circuits. The second category follows the concept of in situ timing error detection, introduced by Ernst et al. with "Razor" [22]. The general idea is the detection of late arriving changes on the data pin in relation to the clock pin on registers identified as potential critical paths. If an error is detected the clock period can either bee stretched out, replaying the previous instruction or the overall frequency can be reduced. Over the years a plethora of variations of the concept have been published [23]–[26].

## 1.3 Body Control for Variability Mitigation and Throughput Scaling

Body biasing as a technique has been around for decades and used to be a fairly effective knob to adjust the threshold even in standard CMOS bulk technologies. Adjusting the body voltage directly effects the threshold voltage $V_T$ as shown in Fig. 1.3 for a simple inverter: when applying a positive voltage to the bulk contact of the NMOS and a voltage smaller than $V_{DD}$ to the bulk contact of the PMOS the circuit is pushed into a forward bias condition. This decreases the threshold voltage $V_T$, resulting in a faster circuit speed and an increased leakage. The other

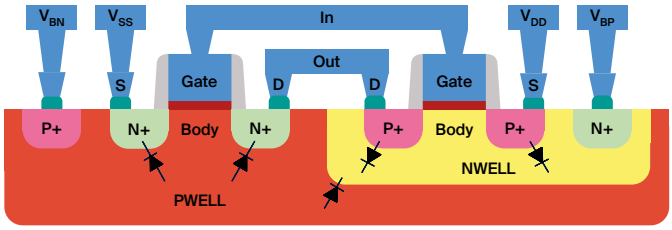



Figure 1.4 – Cross section of the CMOS inverter with the built in diodes.

way around, when applying a negative voltage to the bulk contract of the NMOS and a voltage larger than $V_{DD}$ to the PMOS bulk we achieve a reverse bias with an increased $V_T$ resulting in a slowdown and reduction in leakage.

The limits of the biasing are defined by the built in diodes shown in the cross section of the CMOS inverter depicted in Fig. 1.4. These have to be kept in a reverse bias condition during operation to guarantee functionality and to prevent excessive leakage. The main limitation is the p-n junction formed by the NWELL floating within the PWELL, resulting in a maximum forward bias just below the threshold voltage of the built in diode which is typically around 0.7 V.

More precisely, the threshold voltage change caused by a change in body voltage of the MOSFET can be modelled as follows:

$$V_T = V_{T0} + \gamma \left( \sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|} \right) \tag{1.10}$$

The main contributor here is the body factor $\gamma$, a process specific parameter modelling the effectiveness of changes in the body voltage $V_{SB}$ which in the following will be referred to as $V_{BN}$ and $V_{BP}$ for NMOS and PMOS respectively. $V_{T0}$ expresses the threshold voltage for the zero bias case where $V_{BN}$ is equal to $V_{SS}$ and $V_{BP}$ is at the potential of the supply voltage $V_{DD}$. The remaining parameter $\phi_F$ is the Fermi potential.

Unfortunately, with feature sizes scaled down into the nanometer range that knob became less and less effective $\gamma \approx 0$, to the point where it is typically no longer considered a useful tool for the designers - to the point where early FDSOI technologies completely removed control due to the thick buried oxide with floating body [10, p. 31].

However, in the last decade two technologies have surfaced to counteract this trend - Ultra Thin Body and Box Fully Depleted Silicon on Insulator (UTBB-FDSOI) [27], and Deeply Depleted Silicon on Insulator [28]. Figure 1.5 depicts the general cross section of both devices besides a standard bulk NMOS transistor. Both devices share the common idea to insulate the channel from the bulk. This insulation has two main effects: the electrostatic field becomes more regular and the channel is insulated, removing the bias limitation of the built in diodes, vastly increasing

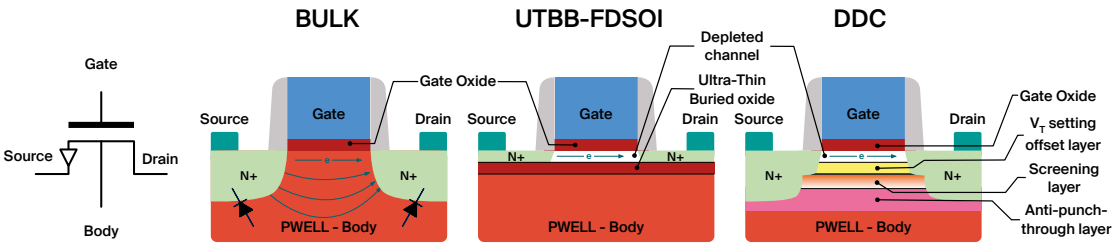



Figure 1.5 – Device cross section of an NMOS transistor in a) a standard bulk technology, b) Ultra Thin Body and Box Fully Depleted Silicon On Insulator [32] and c) Deeply Depleted Channel [30]

the range of the bias.

Furthermore, both technologies share a depleted channel. With nanoscale devices the dopant atoms in the channel become sparse and hence a handful of randomly assigned discrete number of atoms for each device dictates the performance. Hence, if the technology is able to operate on a depleted channel we can expect a significant reduction in random dopant fluctuations which directly translates into a significant reduction of intra die variation, a property which indeed has been shown in literature [29]–[31].

The main difference between UTBB-FDSOI and DDC is the manufacturing process: FDSOI uses SOI-wafers with a thinned down buried oxide layer as a starting point while the later one implants a stack of layers on a conventional bulk wafer in order to form the insulated and depleted channel with tightly controlled threshold voltages.

## 1.4 Body Control with DDC

While UTBB-FDSOI allows for a larger biasing range, its body factor $\gamma$ is more moderate with around 85 $^{mV}/_{V}$ [33], [34] when compared to the 375 $^{mV}/_{V}$ for 55 nm DDC. The process allows the bias to be adjusted in the range of $-1\,V < V_{BB} < 0.6\,V$ which, combined with its particularly large body factor results in a remarkable property: *After* using body biasing for full PVT compensation DDC still allows for a change of drive strength by a factor of roughly 200 and 300 for PMOS and NMOS, respectively [35]. This property opens up a large design space for the designer, to play with frequency, supply voltage, and body control jointly to achieve the best power performance for their particular application. However, finding the best combination of these parameters for a particular design is a non trivial task which will be explored throughout this thesis.

## 1.5 Thesis Outline and Contributions

In this thesis we describe methodologies for design space exploration, exploiting the design space extension through the strong body effect of DDC together with the capabilities of voltage scaling. This thesis is broken down into two major parts:

The first part concentrates on facilitating the muti-dimensional parameter exploration (supply voltage, frequency, body bias) at design time. To this end we focus on standard cell characterisation data to obtain a broad view on the scaling behaviour of a circuit. The intention is to cover the whole variety of different cells as well as transients and load conditions we expect to face in an actual circuit. We then show how to use that data to map the design space in a way which allows for predictions on circuit speed as well as leakage and dynamic power while also providing the necessary bias voltages.

The second part concentrates on actual circuits: Calanda, a test chip designed within CSEM with the goal to prototype an SoC using an on-current based biasing scheme while also integrating some test circuitry to allow for verification of the conclusions from the standard cell characterisation. Nakayama extends the biasing scheme shown in Calanda by integrating a frequency locked loop (FLL) based control on top. Finally, Snaefellsjokull proposes an alternate lightweight direct charge pump based biasing approach using a novel PMOS/NMOS balance sensor for regulation.

### 1.5.1 Part 1: Library Analysis and Design Space Exploration

**Chapter 2, Standard Cell Library Analysis and Comparison**, describes a methodology to characterise the quality of a standard cell library before and after compensation based on library characterisation data. We show that body bias based compensation pushes the PVT corners together into a near match, reducing the cross corner median delay variation by two orders of magnitude.

**Chapter 3, Design Space Evaluation and Low Effort Mapping**, again uses standard cell characterisation data, but this time with the goal of creating a tool for rapid design space exploration. We first identify the design space as the maximum and minimum on currents a transistor can achieve across all process and temperature corners when varying the bias forward and reverse respectively across the supply voltage range. We then sample this design space in equidistant steps by characterising a pruned library which we then use for a 32 bit multiplier as a sample design. The design is then analysed with an STA and power analysis tool in order to extract speed, dynamic power, and leakage at each design point. In the following the data is resampled into a $V_{DD}$ vs. frequency grid and interpolation is used in order to allow for intermediate points. We then show that we can derive universal scaling factors for both leakage and dynamic power, allowing to apply the model based on the reference design to arbitrary circuits, using a single point calibration. On the example of a 32 bit microprocessor we achieve

a near perfect match, predicting the total power within an error margin of one percent of the results of the dynamic and leakage estimation using a specifically characterised full standard cell library at the same points.

### 1.5.2 Part 2: Library Analysis and Design Space Exploration

**Chapter 4, Introduction to Biasing Systems**, gives an introduction into biasing systems, followed by a brief literature review. We break down the general components, and explain concepts and use cases.

**Chapter 5, Calanda: Analog On-Current Regulation**, describes the Calanda SoC. Calanda integrates a 32 bit IcyFlex2 RISC core with 64KB SRAM, 4KB of standard cell memory and some standard microcontroller peripherals such as GPIOs, UART, SPI, and JTAG. The Idea of the on-current based bias control concept designed within CSEM is sketched out and we present power measurements of the core and SRAM under body control. Furthermore, the construction of standard cell based ring oscillators is presented. We apply the bias and supply voltages as predicted by the model introduced in Chapter 3 in an open loop fashion and show well matched constant frequencies across the rings when following a constant frequency trajectory through the design space at different predicted operating points. Furthermore, we show that a single point calibration for the static and dynamic power scaling factors achieves a decent match between the power model and measurements.

**Chapter 6, Nakayama: Analog On-Current Regulation with Secondary FLL based Regulation Loop**, describes the Nakayama SoC. The SoC integrates an icyglex 5, a 32 bit RISC-V-compatible core with 256 KB of bias domain banked SRAM, a BLE PHY, DAC/ADC, Timers, SPI, UART, and GPIOs. Nakayama extends on the biasing concept of Calanda by adding a secondary frequency locked loop (FLL) based control loop, allowing the user to set a target frequency rather than a numerical DAC code. We describe the FLL design and construction in detail and show measurement results.

**Chapter 7, Snaefellsjokull: Worst Case Oriented Body Bias**, describes an alternative approach for a biasing system centred using a light weight concept using a charge pump directly for the well together with a PMOS/NMOS balance system. The design strategy is worst case driven, but we show that the resulting regulation results in decent performance as well for the other cases.

## 1.6 Third-Party Contributions

The Calanda chip on its own was a team effort with multiple persons working on concept, implementation, and integration. The on current based biasing approach, used for both Calanda and Nakayama, has been designed by a large team within CSEM. My role on this chip was mainly on the digital pre study which became Chapter 2 and the concept of the standard cell

based oscillators for hardware verification of the library chapter conclusions. Furthermore, I have driven a large share of the chip characterisation, including the memory design space exploration.

For Nakayama, as for Calanda, the actual biasing system has been designed as a group effort within CSEM. The main contribution of this thesis is in the implementation of the secondary FLL control loop, together with the programmable standard cell based oscillator responsible for generating the circuit clock.

# Part I
# Library Analysis and Design Space Exploration

# 2 Standard Cell Library Analysis and Comparison

In this chapter we characterize the cross corner and cell to cell variation of a custom standard cell library (SCL) in USJC 55 nm Deeply Depleted Channel (DDC) technology. Precharacterized Liberty library files are used as a dataset for comparing both the unbiased library against a characterization using adaptive body biasing (ABB) to compensate for process voltage and temperature (PVT) variation.

The chapter is largely based on the following paper:
T. C. Müller, J.-L. Nagel, M. Pons, *et al.*, "PVT compensation in Mie Fujitsu 55 nm DDC: A standard-cell library based comparison", in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2017, pp. 1–2. DOI: 10.1109/S3S.2017.8309246,

## 2.1 Introduction

When implementing an adaptive PVT compensation scheme, the designer is interested in some measure of the effectiveness of the chosen approach. Typically this measure is derived from a simple model of inverters or ring oscillators which is easy to integrate into the workflow typically used by analog designers. The assumption is, that these simple models, often based only on inverters, capture the variability across corners and the impact of changing the bias voltage by some common factor across all corners and for all instances of the digital cells within the circuit.

However, complex digital circuits typically contain a wide spread of cells, instantiated in different load scenarios which will see a wide range of input transitions. This adds additional uncertainty, not following a common factor, potentially resulting in critical path shifts across corners. This complicates multi-corner design-time optimisation and increases margins for run-time PVT compensation.

For the SoC designer it is important to know whether the chosen method of compensation applies identically to all the cells in the employed cell library or whether he needs to apply

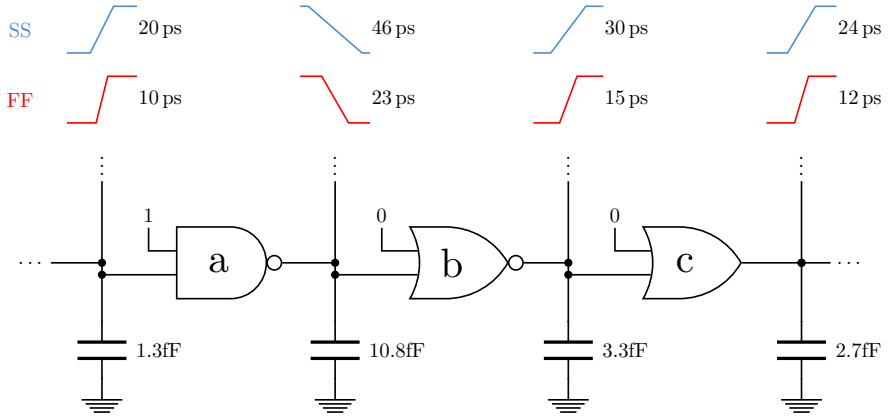



Figure 2.1 – Transitions seen in a sample segment of a critical path in two operating corners (blue/red), arbitrary numbers for illustration purpose only.

margin to account for the delay scaling differences between cells during the implementation. If so, finding the right amount of required margin needed is a non trivial task.

Furthermore, designers of a standard cell libraries are interested in an exhaustive sweep of all possible conditions to identify the cases which cope particularly badly with variation. Setting up these kind of simulations in an analog design environment is a tedious task and it is easy to miss cases.

Digital implementation tools however already deal with a similar problem: they need transition and delay data in order to be able to provide an accurate estimate of static timing for the segments of the critical path as illustrated in Fig. 2.1. Hence, characterisation tools have been utilised for years to automatically generate test harnesses around cells in order to characterise the cells running exhaustive spice simulations in order to generate the necessary data.

In the following we will exploit the data generated from such a characterisation tool to extract the efficiency of an adaptive PVT compensation scheme.

## 2.2 Cross Corner Library Comparison Methodology

When analyzing a circuit across operating corners we observe a scaling of the achieveable clock frequency $f_{\max}$. In first order approximation this scaling can be described as a factor $\alpha^l$ for an operating corner $\text{þ}_l$ relative to a reference corner $\text{þ}_{\text{ref}}$.

$$\alpha^l = \frac{f_{\max}(\text{þ}_{\text{ref}})}{f_{\max}(\text{þ}_l)} \tag{2.1}$$

The actual limiting factor however is the accumulated delay within the most critical path of the actual circuit, built from a specific selection of standard cells. Each cell $c$ is constructed with a different schematic, resulting in minute differences of cross corner scaling. To provide a more

specific characterisation compared to (2.1) we could derive a naïve per cell scaling factor

$$\alpha_c^l = \frac{t_{\mathrm{pd}}(þ_l)}{t_{\mathrm{pd}}(þ_{\mathrm{ref}})} \tag{2.2}$$

based on the cell propagation delay $t_{\mathrm{pd}}$. This approach however is still too general to be meaningful in the context of an actual circuit. First, when transitioning from one logic input configuration to another one, different parts of the cell schematic will be excited based on the inputs, causing different delays. The most obvious cases are the pull up and pull down networks being responsible for rise and fall transitions respectively. The second, and more severe factor, is the actual environment seen by the cell, defined through the load on the output as well as the steepness of the transition on the input.

### 2.2.1 Identifying Identical Operating Points Between Two Corners

When comparing the timing (i.e. instance delays and transition times) of a circuit for different PVT conditions we see a significant difference not only in the delay, but also in the transition times determined by the driver of each individual cell as illustrated in Fig. 2.1. Hence, we can not simply use the delay scaling for a given transition time of the individual cells as a proxy for the delay of the overall circuit since the parallel transition time degradation is not reflected properly.

As we would like to extrapolate the circuit delay scaling, including the impact of degraded input transitions, from the delay characteristics of individual cells in the library, we need to compare timing arcs that result from different input transitions.

To this end, we first derive characteristic transition times for each load $C_{in}$ as specified in the library for each considered PVT corner. This characteristic transition time for a given load is obtained from a reference gate—for the purpose of simplicity a buffer—that is replicated in a chain in which each node is loaded with $C_{in}$ as illustrated in Fig. 2.2. Any reasonable transition applied on the input of the chain quickly converges to a characteristic transition time determined by the load $\mathrm{C_{in}}$ within a few steps. The implementation is a simple iterative table lookup with interpolation until conversion is reached as shown in Alg. 1.

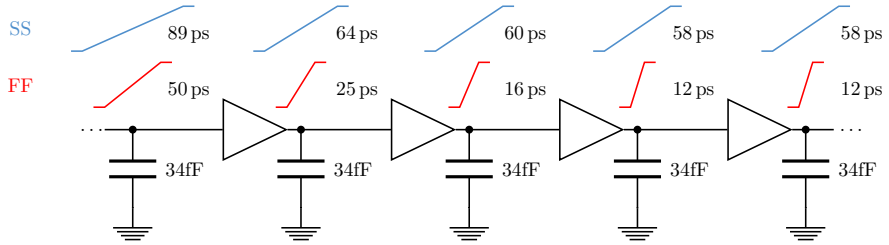



Figure 2.2 – The transition at the output of the chain is characteristic for the combination of load (upper vs. lower chain) as well as operating corner (red vs. blue). The high gain of CMOS process results in a quick recovery of slow transition through a path and thus the output transition can be considered independent from the input transition for a long enough chain. Arbitrary numbers for illustration purpose only.

**Algorithm 1** Iterative derivation of the characteristic transition for a given load.

```
function ExtractTransition(cell,load,corner)
    currentT ← centerTransition(cell, load, corner);
    while iteration < iterationLimit do
        nextT ← transition(currentT,cell, load, corner);
        if |nextT − currentT| < ϵ then;
            return nextT;
        else
            currentT ← nextT;
        end if
    end while
    return currentT;
end function
```

We now compare cells and their timing arcs in different PVT corners no longer for the same fixed transition time, but for the corresponding characteristic transition times within the respective corners associated with the same—PVT independent—load. With this approach cell delays are compared based on the equivalent (not equal) transition times in both corners.

### 2.2.2 Comparing Individual Timing Arcs

In the following we refer to a specific tuple of a cell $c$, an input configuration $p$, a load $C_{\text{load}}$, and an input transition $t_{\text{rf,in}}$ within a standard cell library as a realization $r = (c, p, t_{\text{rf,in}}, C_{\text{Load}})$. Thus, we can define a realization specific delay scaling factor $\alpha_r^l$.

$$\alpha_r^l = \frac{t_{\text{pd}}(r, \text{þ}_l)}{t_{\text{pd}}(r, \text{þ}_{\text{ref}})} \tag{2.3}$$

When comparing the values $\alpha_r^l$ derived from the library across corners without compensation we expect a large spread, reducing the $t_{\mathrm{pd}}$ for either fast process corner, low temperature or higher supply voltage while slow corner, high temperature or lower supply voltage cause an increased $t_{\mathrm{pd}}$ and thus larger $\alpha_r^l$. Comparing compensated libraries on the other hand, values of $\alpha_r^l$ are expected to be found close to one, where the spread around one is a measure for the effectiveness of the compensation (smaller spread is better).

## 2.3 Dataset

The library analysis has been conducted on a custom standard cell library with 69 cells. It was designed for near threshold operation in USJC 55 nm DDC technology utilizing BB for PVT compensation. The library utilises 90nm channels on ultra low leakage (ULL) transistors in order to reduce the leakage even further and implements PMOS and NMOS transistors of identical size, relying on body control to balance the corresponding drive strengths.

For the uncompensated baseline data set we apply VDD and VSS to the PMOS and NMOS bulk contact respectively. For the compensated data set the the bias voltages have been derived using Spice simulations, setting the bias voltages such that we achieve identical on-currents for NMOS and PMOS across the corners. The underlying assumption was that constant on-currents result in identical speed across the corners.

The standard cells have then been characterised for the selected operating corners using Cadence Liberate. The analysis was done on the output files in the Liberty file format, using the non linear delay model (NLDM).

We decided to use the typical process corner at 25 °C and 500 mV with some forward biasing as reference corner $\mathfrak{p}_{\mathrm{ref}}$ for normalization of both the biased and unbiased case to allow for direct comparison.

## 2.4 Results

Figure 2.3 shows the necessity of the use of characteristic transitions in order to achieve a fair comparison between corners: the box plots show the variation seen for a corner which is far away from the reference corner as well as for an other one which is fairly close. While the use of characteristic transitions has very little effect on the close corner (median nearly identical, reasonably close match of the upper and lower quartile) the effect is significant on the much slower corner: with the naïve direct comparison of identical input transitions we obtain an overly pessimistic prediction of the factor alpha with a significantly larger spread.

The boxplots in Fig. 2.4 show the distribution of the $\alpha_r^l$ within each operating corner, before (top) and after compensation (bottom). SS, 450 mV, −40 °C is an extreme case where the circuit drops into deep subthreshold operation if not properly compensated through the biasing system.

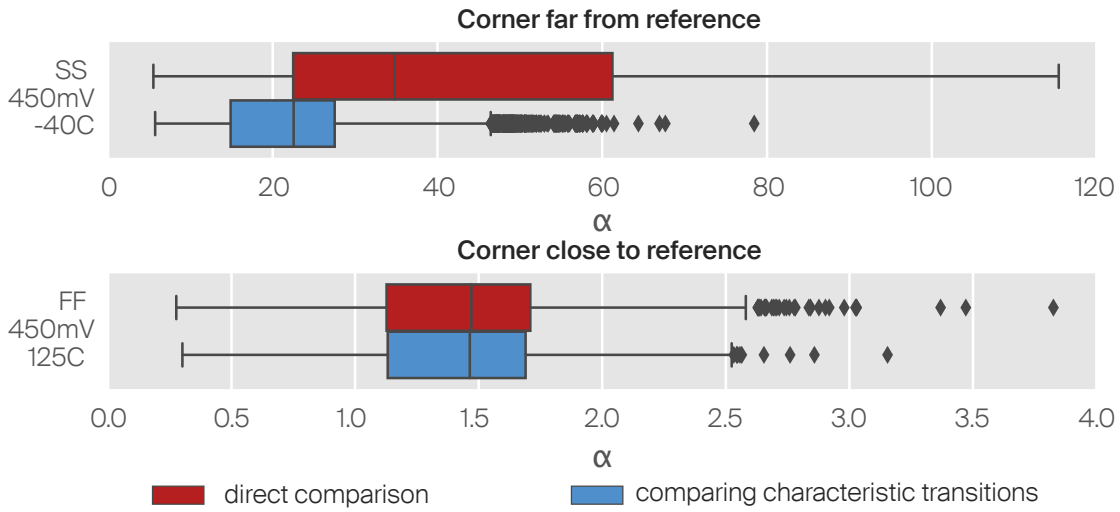



Figure 2.3 – Without the use of characteristic transitions to obtain equivalent load based transitions the corners far from the reference corner show an overestimated variation, while the effect vanishes the closer the corners get to each other.

A worst case timing estimate that considers all the outliers (marked with the black crosses) would result in an overly pessimistic setup margin of 7845 % for the unbiased case. With the application of ABB this pessimistic margin shrinks to 431 %.

Since these margins are clearly excessive, we next analyze the responsible outliers to remove those that are not relevant in a practical circuit. Figure 2.5 plots the values of $\alpha_r^l$, sorted in descending order with and without compensation. We observe a steep drop on the left side of both plots, suggesting that only a small number of synthetic realizations are responsible for the majority of the observed worst-case delay variation. For an actual path within a design we would, however, expect only a few of these outliers if any at all, limiting their potential influence on the overall timing. Thus, for the majority of the designs these outliers can be ignored when discussing practically relevant margins.

We further note that a large share of the outliers are due to extreme input slope and load conditions which are typically prohibited by the design constraints used during implementation: it is therefore unlikely to see for example a minimum sized inverter driving high load net and hence those corner cases can be excluded from the analysis.

Instead, considering the wide plateau of similar values of $\alpha_r^l$ around the center of Fig. 2.5, a more reasonable estimate for the cross corner performance of a circuit is the median $\tilde{\alpha}^l$, marked blue in Fig. 2.4. With compensation this median is pushed from a range of $-9\%$ to $2154\%$ towards $-18\%$ to $18\%$, effectively reducing the margins needed to cover cross corner variation.

In order to observe the per corner, cell to cell variation we can remove the cross corner effects

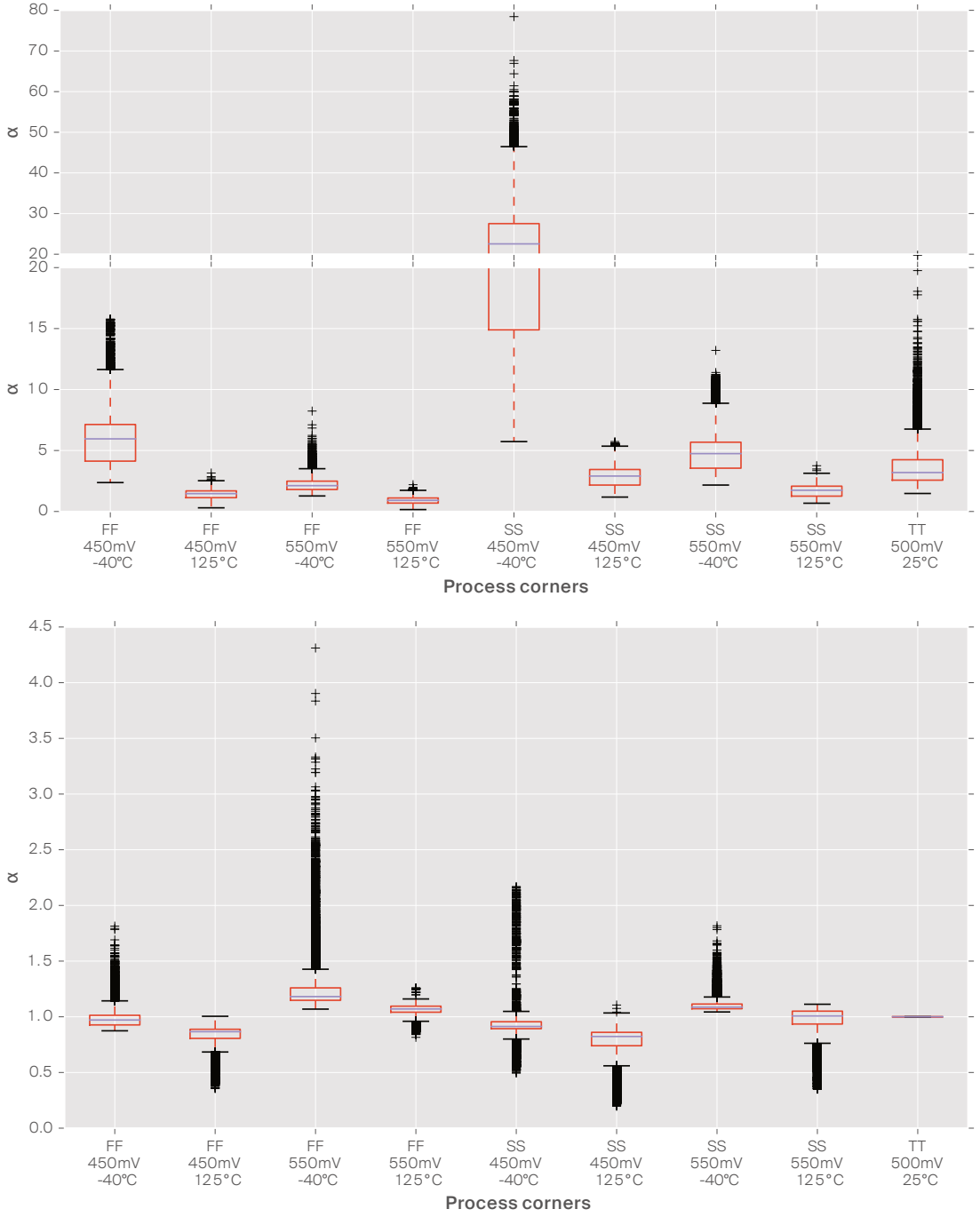



Figure 2.4 – Boxplots of $\alpha_r^l$ without (top) and with (bottom) compensation. To improve the visualisation the axis has been split at an $\alpha_r^l$ of 20, showing the upper remainder of the box with a compressed axis.

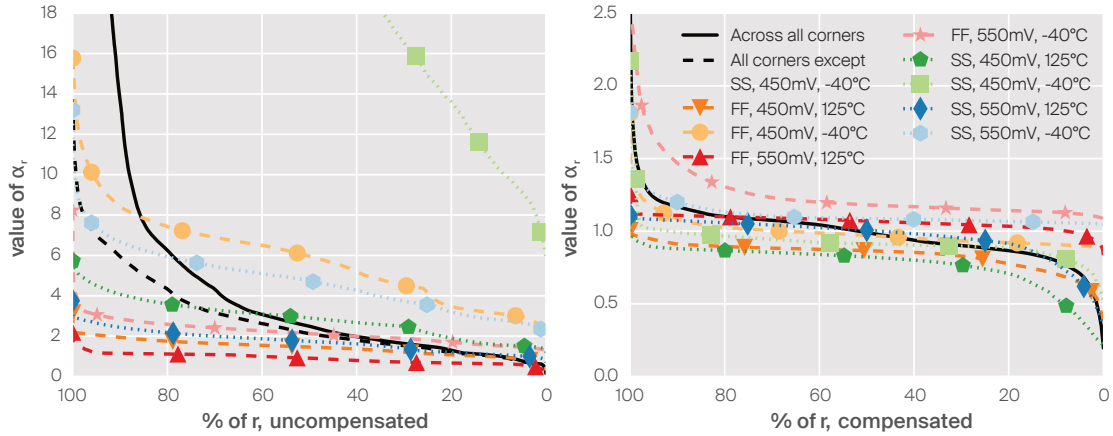



Figure 2.5 – Sorted $\alpha_r$, ordered from slowest to fastest.

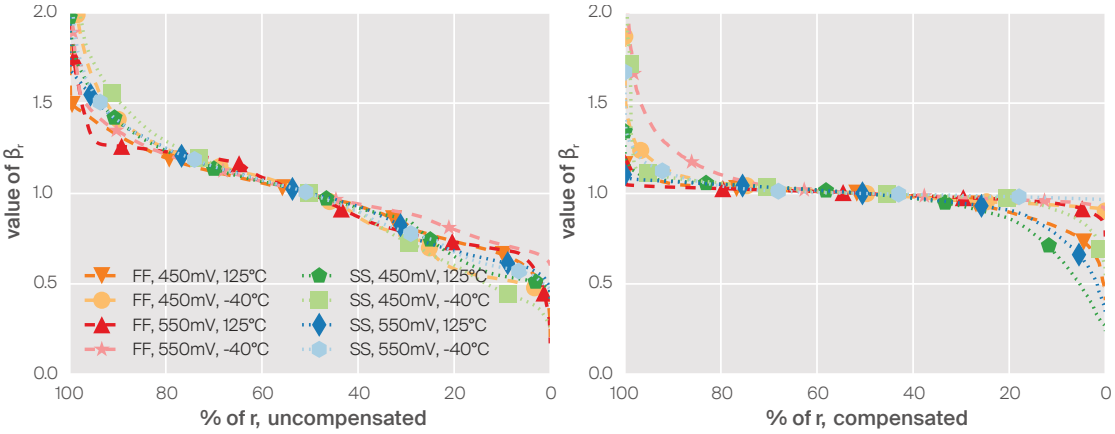



Figure 2.6 – Sorted $\beta_r$, ordered from slowest to fastest.

by normalizing each $\alpha_r^l$ with the median $\tilde{\alpha}^l$ to obtain a within corner scaling factor $\beta_r^l$

$$\beta_r^l = \frac{\alpha_r^l}{\tilde{\alpha}^l}, \tag{2.4}$$

that reflects the variation across realizations.

The result of the normalization is depicted in Fig. 2.6. We observe that all corners are pushed on top of each other both for the compensated and uncompensated case. This observation shows that the median tracks the global scaling factor very well. Furthermore, we observe a pronounced plateau after compensation, implying even a reduced within-corner variation.

## 2.5 Conclusion

We were able to exploit standard cell characterisation data in order to compare arbitrary corners against each other for a comprehensive set of timing arcs, transition times and load

combinations. We proposed a methodology for a fair "apples-to-apples" comparison of synthetic circuit realizations in order to get a broad idea of the effectiveness of a compensation approach. On the example of a low power standard cell library designed for 55 nm DDC using adaptive body biasing, we showed that ABB is capable of compensating the PVT variation, achieving a reduction of the cross corner median delay variation by two orders of magnitude while also achieving a reduction of per-corner cell to cell variation.

# 3 Design Space Evaluation and Low Effort Mapping

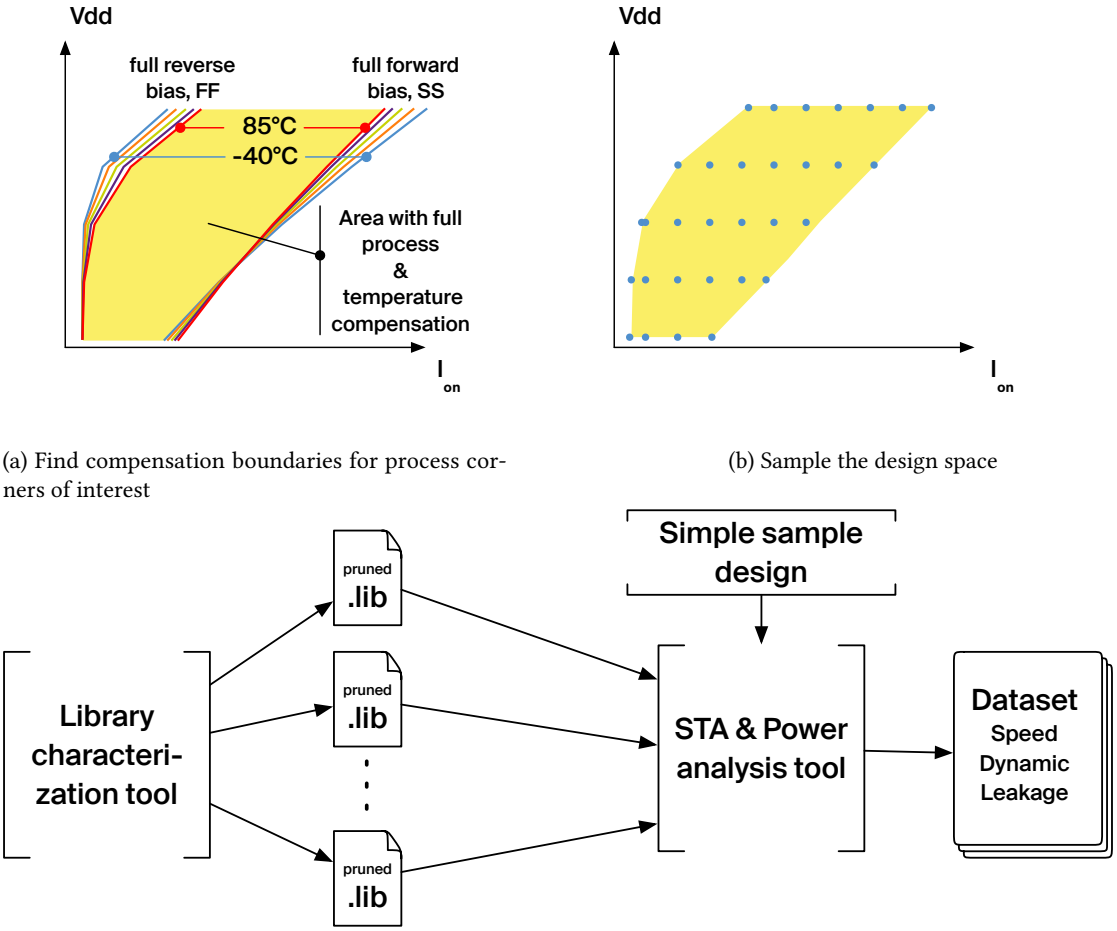
In this chapter a systematic low-power design methodology for technologies that offer a strong body factor is discussed. Specifically, we explore both the body bias voltage and the supply voltage knobs in order to find the MEP (minimum energy point) for a constant target frequency.

The methodology presented accounts for process and temperature (PT) variations while charting the design space for a simple reference design. We then show how to scale the energy data of this reference design to any arbitrary design. A case study of a 32 bit RISC microprocessor achieves an energy estimation match of our significantly less complex estimation methodology within 1% of traditional signoff results.

This chapter is based on the following publication:
C. T. Müller, M. Pons, D. Ruffieux, *et al.*, "Minimum Energy Point in Constant Frequency Designs under Adaptive Supply Voltage and Body Bias Adjustment in 55 nm DDC", in *2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, Jul. 2019, pp. 285–288. DOI: 10.1109/PRIME.2019.8787736

## 3.1 Introduction

Technologies with a strong body factor such as USJC 55 nm deeply depleted channel (DDC) [29] provide the designer with an additional knob for energy optimisation in addition to the widely used adaptive voltage and frequency scaling (AVFS) techniques for which an analytical energy model was presented [38]. The goal of the designer is to operate as close as possible to the operating point where the sum of the leakage energy and the dynamic energy is minimal (MEP). Typically this MEP is located within sub- or near threshold operation, which significantly reduces the maximum operating frequency [14], [39]. However, with the body voltage as an additional knob, adaptive supply and body bias scaling (ASVBB) can be implemented, allowing for MEP tracking at a given constant frequency which is determined only by the application requirements [40]. Recently [41] presented such a scheme to automatically track the MEP at a given frequency by pinning the ratio of leakage to dynamic energy into a predefined range. In




(a) Find compensation boundaries for process corners of interest

(b) Sample the design space

(c) Characterize a pruned library across the design space.

(d) Extract reference design speed, dynamic power and leakage.

Figure 3.1 – Steps to extract the reference data over the design space.

light of the efficiency of this approach and the need for real-time systems that require a given frequency, there is a need for a methodology to rapidly explore the associated design. Such a methodology rapidly identifies the achievable limits and provides near-optimal settings for adjusting supply voltage and body bias.

## 3.2 Mapping the Design Space

We start by charting the ASVBB design space, which defines the spread of leakage, dynamic energy, and frequency operating points.

Using ASVBB we may vary the supply voltage with the lower limit set by the reliability of the circuit and the upper limit set by the process. The legal body bias range $\mathcal{V}(V_{dd})$ is set by the supply voltage as well as technology parameters, most notably the structure and forward

voltages of the built-in body diodes between NWELL and PWELL. Furthermore, the valid range under nominal conditions is reduced when PT variations have to be compensated with legal settings. For the USJC 55 nm DDC process used in this work, the strong body factor of 375 mV/V, allows to fully compensate across PT while still retaining a large frequency design space for MEP tracking [35].

For the purpose of ASVBB design space exploration we use the on-current $I_{ON}$ as a proxy for frequency, based on the following first order aproximation where $f$ denotes the frequency, $C$ the load capacitance, and $V_{dd}$ the supply voltage.

$$f \propto \frac{I_{ON}}{C \cdot V_{dd}} \tag{3.1}$$

Contrary to the frequency, $I_{ON}$ can easily be measured and simulated while the PMOS and NMOS strength can be trivially fixed to a constant ratio. Furthermore the $I_{ON}$ can be kept constant across PT, as the adjustment of the threshold voltage through the body voltage counteracts the effects of variation. The design space limits are then determined through spice simulations of $I_{ON}$ through a representatively sized transistor under application of the maximum forward ($V_{bb}^{fwd} = 0.6\,\mathrm{V}$) and reverse ($V_{bb}^{rev} = -1\,\mathrm{V}$) bias voltages, for the expected operation temperatures $T \in \{-40\,^{\circ}\mathrm{C}, 25\,^{\circ}\mathrm{C}, 85\,^{\circ}\mathrm{C}\}$, and across the process corners $P \in \{FF, SS, TT\}$ of interest. The low end of the design space is defined by the highest $I_{ON}$ for which PT can still be compensated (i.e., maximum reverse bias):

$$I_{ON}^{min}(V_{dd}) = \max_{p \in P, t \in T} \min_{V_{bb} \in \mathcal{V}(V_{dd})} I_{ON}\left(V_{dd}, V_{bb}, p, t\right) \tag{3.2}$$

Similarly, the upper end of the design space is defined by the minimum $I_{ON}$ for which PT can still be compensated (i.e., maximum forward bias):

$$I_{ON}^{max}(V_{dd}) = \min_{p \in P, t \in T} \max_{V_{bb} \in \mathcal{V}(V_{dd})} I_{ON}\left(V_{dd}, V_{bb}, p, t\right) \tag{3.3}$$

Figure 3.1a shows the boundaries set by the sweep, with the ASVBB design space spanned between the minimum $I_{ON}$ (3.2) and the maximum (3.3) for each supply voltage. We notice that the design space across corners is very similar to the design space of the individual corners, thanks to the large body factor.

## 3.3 Reference circuit

After defining the boundaries of the ASVBB design space, we proceed to characterize normalized active energy, leakage power, and timing within this valid range. As shown in Fig. 3.1b, we propose to sample the ASVBB design space with a grid defined by sweeping a) the supply voltage and b) $I_{ON}$ from minimum to maximum. For each grid point PMOS/NMOS bias voltage pairs are obtained. With these bias voltage pairs a standard cell library can be characterized for

each grid-point and for each PT combination. By using these libraries, we can easily search the design space using power and timing analysis for the optimum operating points of the target design.

### 3.3.1 Standard Cell Library Characterisation

Unfortunately, a full standard cell library characterisation for many operating points is often prohibitive due to long simulation times, in particular if a dense grid is expected for high precision. We therefore propose to limit the analysis to only a few cells, representing a distribution of cells commonly found in a large design. The most significant reduction in characterisation time can be achieved by dropping sequential cells from the library. These need to be characterised for setup and hold constraints which comes with the need of extensive sweeps. With non-sequential cells accountable for the majority of the delay in the critical path of a typical design, the effect of potential scaling differences between sequentials and non sequentials can be considered negligible.

To select a representative subset of cells, a purely combinatoric 32 bit multiplier was synthesized with RTL compiler, constrained to a maximum delay of zero from all inputs to all outputs. The cell distribution was extracted, rarely used cells where pruned, and the circuit was resynthesized using only the top thirteen cells, which cover over 90% of the cells in the original design. This pruning approach reduced the simulation time of our characterisations from slightly short of two hours to less than three minutes, allowing to characterize 55 points across three process corners and five temperatures on a single machine over a weekend. Finally, leakage, dynamic power as well as the frequency can be extracted for the reference design after re-synthesis with only the selected cells.

### 3.3.2 MEP for the 32 Bit Multiplier Reference Design

Using interpolation, we can now approximate the dynamic and leakage power across the whole ASVBB design space, which allows us to find the MEP for the operating modes of interest. Figure 3.2 shows the tradeoffs for the multiplier reference design, tracing the minimum energy point across the range of target frequencies from 20 MHz to 140 MHz. On the low end, the minimum supply voltage of our design space is the limiting factor, with the bias as far reverse as possible to still reach the frequency. Subsequently only the bias increases until an inflection point is reached around 50 MHz where the supply voltage starts to increase, causing an inversion of the bias effectively reducing the leakage. At the 80 MHz frequency cut, the typical MEP curve is formed, with the higher forward bias at low supply voltage causing the leakage energy to dominate below 0.6 V and the dynamic energy dominating at a higher supply. At the 120 MHz frequency cut, the design space is limited again, with the speed not being reachable at a supply voltage below 0.7 V.

Similarly, Fig. 3.3 shows the shift of the MEP across frequency for different process corners,

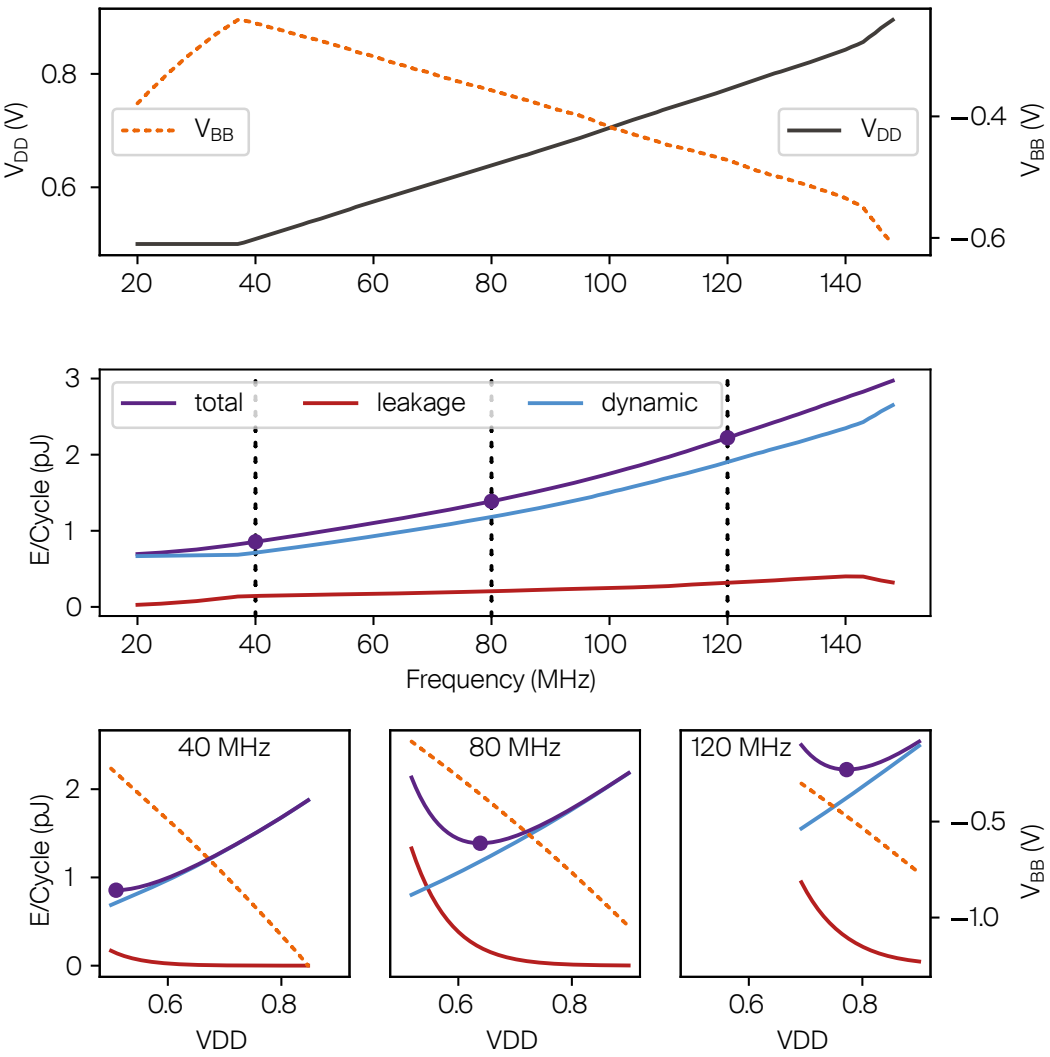



Figure 3.2 – Top: The bias and supply voltage pairs for minimum energy across frequency at TT/25 °C. Center: The corresponding leakage and dynamic Energy components. Bottom: minimum energy curves for the three marked frequency cuts in the center plot.

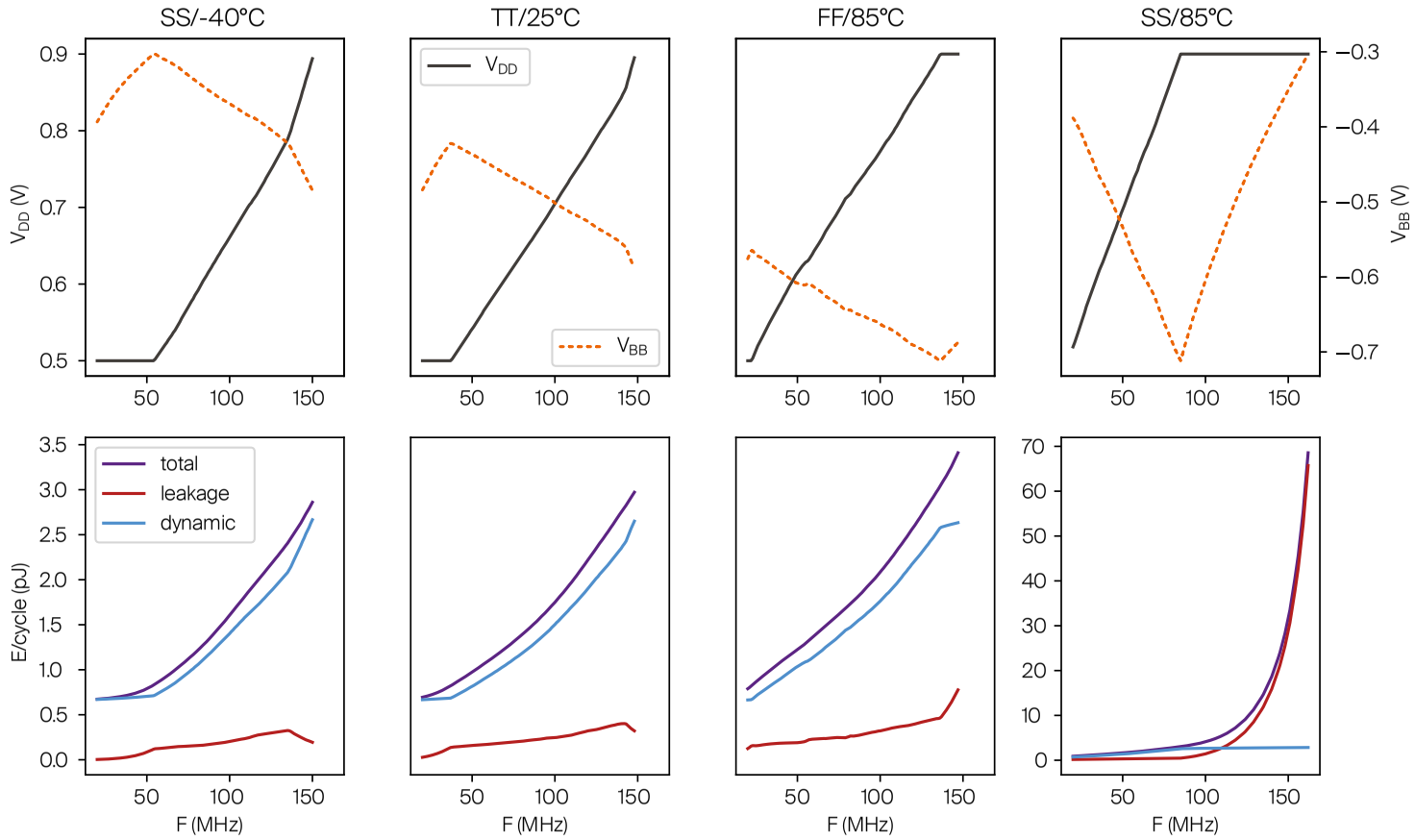



Figure 3.3 – Top: Supply and bias voltages at the MEP for different target frequencies in four process corners. Bottom: Energy across frequency at these points, broken down into the leakage and dynamic components.

when adjusting both bias and supply voltage for minimum energy at each frequency step. Note that, as long as a tradeoff exists, the absolute energy values for MEP are almost constant over PT for constant frequency, as the same $I_{ON}$ results in the same dynamic energy while the leakage is kept under control. However, when the edge of the design space is reached as illustrated by the SS/85 °C case in Fig. 3.3 only biasing remains for frequency adjustment causing the leakage to explode.

## 3.4 Scaling the Reference Circuit to an Arbitrary Design

Obviously, the MEP and the associated optimal supply voltage and body bias are highly design dependent since the share between leakage and dynamic energy is determined by the activity and the critical path length [39]. Hence the analysis has to be done for each design as well as for each utilisation profile. We propose to use the small reference design with a limited number of cells to calibrate a model that can then be used to extract the MEP and the associated ASVBB settings for different target frequencies. In the following we discuss the corresponding scaling model.

### 3.4.1 ADVBB Model

The per cycle energy of a design can be split into the sum of the dynamic contribution $E_{dyn}$ and the leakage contribution $E_{leak}$ so that $E_{tot} = E_{dyn} + E_{leak}$. The two energy components are obtained with the following first order approximations in which $\alpha_{sw}$ denotes the design specific switching activity, $C_L$ the total load capacitance, $I_{OFF}$ the leakage current, $V_{DD}$ the supply voltage, and $t_p$ the cycle time:

$$E_{dyn} = \alpha_{sw} \cdot C_L \cdot V_{DD}^2 \tag{3.4}$$

$$E_{leak} = V_{DD} \cdot I_{OFF} \cdot t_p. \tag{3.5}$$

The dynamic energy ratio $r_{dyn}$ between the dynamic energy of the target design $E_{dyn}^{target}$ and the dynamic energy of the reference design design $E_{dyn}^{ref}$ is defined as

$$r_{dyn} = \frac{E_{dyn}^{target}}{E_{dyn}^{ref}} = \frac{\alpha_{sw}^{target} \cdot C_{tot}^{target}}{\alpha_{sw}^{ref} \cdot C_{tot}^{ref}}, \tag{3.6}$$

where $\alpha_{sw}^{target}$, and $C_{tot}^{target}$ as well as $\alpha_{sw}^{ref}$, and $C_{tot}^{ref}$ denote the load and activity of the target and reference design, respectively.

Similarly, we define the leakage energy ratio between the leakage of the target design $E_{target}^{ref}$

and the leakage of the reference design $E_{leak}^{ref}$ as

$$r_{leak} = \frac{E_{leak}^{target}}{E_{leak}^{ref}} = \frac{g_{target} \cdot \overline{I_{OFF}^{target}} \cdot t_p^{target}}{g_{ref} \cdot \overline{I_{OFF}^{ref}} \cdot t_p^{ref}} \tag{3.7}$$

The precise off currents $I_{OFF}^{target}$ and $I_{OFF}^{ref}$ depend on the threshold voltage $V_t$ which is a function of $V_{bb}$ and the specific schematics of the gates used. However, the off currents of larger circuits can be approximated well by the number of gates $g_{target}$ and $g_{ref}$ and the average cell leakage currents $\overline{I_{OFF}^{target}}$ and $\overline{I_{OFF}^{target}}$ in the two designs. In fact, our results will show that with a sufficiently representative reference design, with reasonable constraints and a comparable cell distribution to the target design we can assume $\overline{I_{OFF}^{target}} \approx \overline{I_{OFF}^{ref}}$ across the full range of supply and bias conditions to justify

$$r_{leak} \quad = \quad \frac{g_{target} \cdot t_p^{target}}{g_{ref} \cdot t_p^{ref}}. \tag{3.8}$$

While the *minimum* cycle times $t_p^{target}$ and $t_p^{ref}$ in (3.8) depend on $V_{DD}$ and $V_{bb}$, we note that we are only interested in optimizing for a given cycle time $t_p^{target}$ for the target design. Further, the cycle time $t_p^{ref}$ for the reference design at the characterization points is anyway known. Hence, the scaling of the reference design in (3.7) and (3.8) becomes independent from $V_{DD}$ and $V_{bb}$ which is the foundation for the subsequent target-design specific optimization.

### 3.4.2 Finding the MEP for a Constant Frequency

To scale the reference design dynamic and leakage characteristics to the target design, the latter is first synthesized at any valid $V_{DD}$ and $V_{bb}$ combination for the desired target frequency. At this point we then extract its dynamic and the leakage energy and compute the corresponding scaling factors $r_{dyn}$ and $r_{leak}$ relative to the reference design at the same supply and bias voltages. With $r_{dyn}$ and $r_{leak}$ only depending on design specific properties, they can be used to directly scale the dynamic and leakage energy from the reference design for any other point in the iso-frequency design space.

$$\begin{aligned} E_{target}(V_{dd}, V_{bias}) \approx\ & r_{leak} \cdot E_{leak}^{ref}(V_{dd}, V_{bias}) \\ & + r_{dyn} \cdot E_{dyn}^{ref}(V_{dd}, V_{bias}) \end{aligned} \tag{3.9}$$

A straightforward exhaustive search based on (interpolated) characterization data from the reference design yields the MEP $\min\{E_{target}(V_{DD}, V_{bb})\}$ and the corresponding voltages.

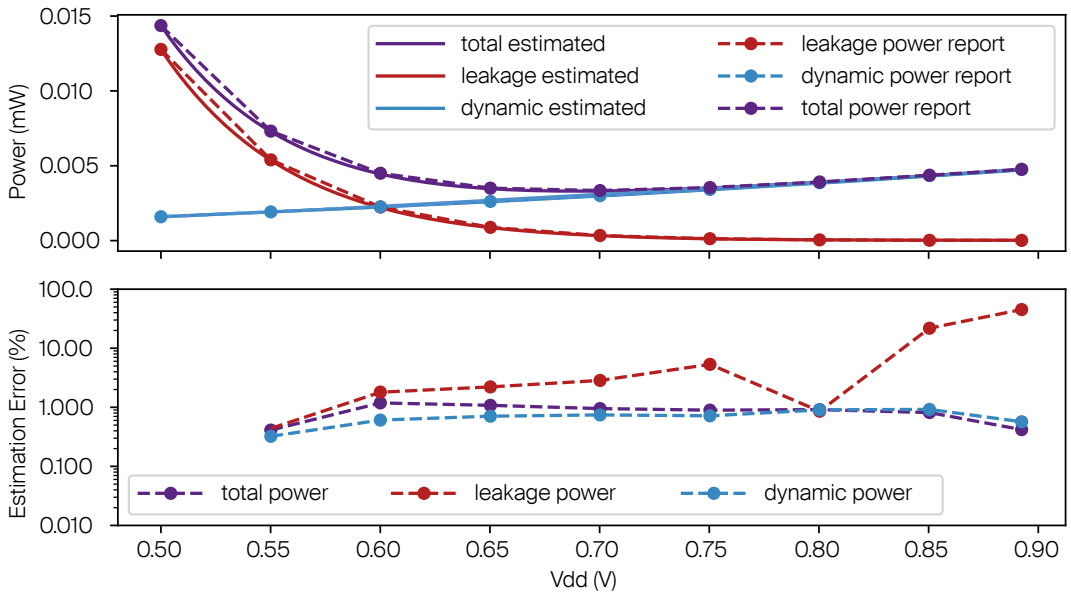


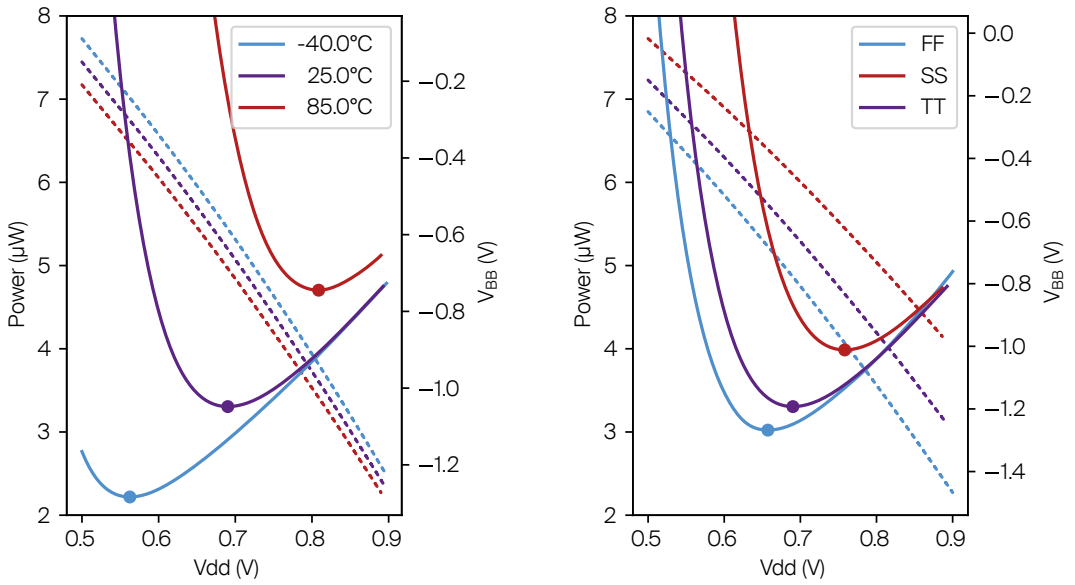

Figure 3.4 – 32 Bit RISC processor: Constant frequency power estimation based on the 32 bit multiplier reference design vs. power report of the design at a given point (circle).

## 3.5 Case Study: 32 Bit Microprocessor

Our target design is a 32 bit RISC microprocessor, synthesized for a relaxed clock of 8 MHz in order to allow for a sufficiently large design space.

### 3.5.1 Modelling Accuracy

First, we evaluate the accuracy of the scaling approach. The microprocessor is implemented at 0.5 V with forward bias to achieve the target frequency. The leakage and dynamic power are extracted and the scaling factors are derived. At constant frequency, the supply/bias tradeoff power curve is then sampled in 50 mV steps at which the full library is charcterized in order to allow for power reports. This curve is shown in Fig. 3.4. The circles denote the energy derived from the power reports and the contineous solid lines denote the estimate based on the scaled ASVBB model. The error in terms of total energy is on the order of 1% over the full voltage range (for a constant frequency). Only the error on leakage increases as the design shifts towards higher supply voltages. However, in these operating points, the leakage impact on the total power becomes negligible as dynamic energy dominates the total power. The leakage mismatch is likely due to numeric effects or lack of precession when reporting.




(a) Across temperature, TT corner

(b) Across process, 25 °C

Figure 3.5 – Power estimation for the 32 bit RISC processor when sweeping the supply voltage across a constant frequency trajectory, trading of supply voltage against bias. Solid: power curves. Dashed: the corresponding body bias $V_{BB}$.

### 3.5.2 Process and Temperature Effects

With the verification of the modelling accuracy we can now use this model to analyze the designs tradeoffs across PT. Figure 3.5a shows the effect of temperature on the MEP in the typical corner: we observe a range of the optimum supply voltage from around 560 mV to 690 mV and 810 mV for −40 °C, 25 °C, and 85 °C, respectively due to the leakage penalty of the forward bias needed to achieve the operating frequency at low supply voltages. A similar picture, however less severe, appears when analyzing the range across process corners as shown for 25 °C in Fig. 3.5b.

## 3.6 Conclusion

The design space mapping methodology presented in this chapter allows a designer to quickly evaluate the tradeoffs of the body bias voltage against the supply voltage for constant frequency operation. We show how to apply precharacterized leakage and dynamic energy maps from a reference design with a pruned library to a larger design with a different critical path length and a different distribution between dynamic energy and leakage power. The accuracy of this model is on the order of 1% of the energy report from the signoff tool. Further, we show how the minimum energy point can shift over a wide range when taking temperature and process variation into account. Joint adjustment of both body bias and supply voltage is key to achieve

the best results.

# Circuits for Bias Control Part II

# 4 Introduction to Biasing Systems

This chapter is intended as an introduction into biasing systems, providing an overview over potential use-cases and their implications with regard to control complexity and regulation duty cycle. We follow up with a quick overview of the components of a biasing system, their use and common problems. We then provide a brief literature review, giving an overview over the state of the art before finally finishing off with an overview over the three biasing systems designed for USJC 55 nm discussed in the following chapters of this thesis.

## 4.1 Use Cases of Biasing Systems

Body biasing modulates the threshold voltage $V_T$ of the MOSFET transistors by applying a bias voltage on the back into the well of the transistor. It is capable of compensating for variations impacting directly the threshold voltage, such as global process variations, ageing, and temperature as well as for fluctuations of the supply voltage. If a process imposes a strong body factor it can be effectively exploited by the designer for all kinds of use cases, as presented in Fig. 4.1. As shown these applications have different requirements in terms of control speed and complexity. In general the use cases can be split into two main directions: On one hand, compensation for the effects of changes in environment and process parameters have on the circuit and on the other hand the application driven goal of dynamically trading power against circuit speed. Lets now have a deeper look on the use cases sketched in Fig. 4.1:

First, the die-to-die process variation are caused by fluctuations present in in dopant concentrations due to small variations during manufacturing leading to a shift in the threshold voltage. These parameters are typically fixed for a given instance of the chip, but may vary noticeably across the wafer and even within the die for large chips. Typically we can assume these to be static throughout the lifetime of the chip, hence control can be achieved statically with a one-time calibration. Body biasing can be applied to compensate by shifting the threshold back towards the expected value.

As circuits are operated an ageing effect will manifest slowly over time. MOSFETs in particular

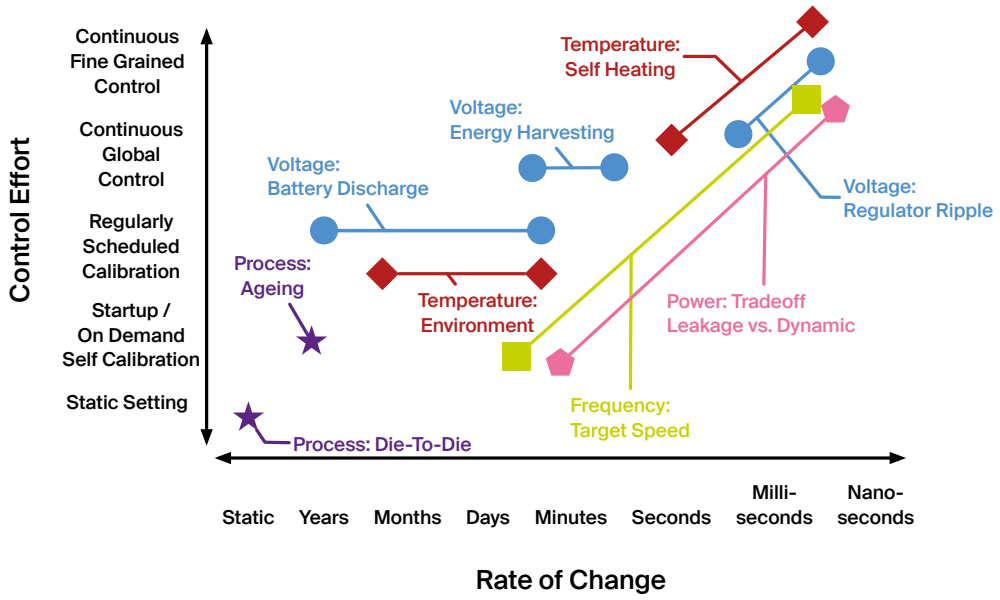



Figure 4.1 – Potential use cases for biasing systems: Compensation for process, voltage and temperature effects as well as setting a target for the circuit speed against power tradeoff.

are prone to two effects which change the threshold voltage over a long time: bias temperature instability (BTI) is an effect on the gate, accumulating charges with voltage stresses due to the potential seen across the gate dielectric. The second one is hot varrier injection (HCI), degrading the mobility and threshold voltage due to damages caused by accelerated carriers traversing the channel during operation [42]. These processes result in a slow degredation of the threshold voltage which can be covered by body biasing calibrating the system with large duty cycle with some form of BIST.

Second, the supply voltage can change over time. As we have seen in Chapter 1 a small change in supply voltage does not have a major effect on the threshold voltage, but on the drive strength of the transistors and hence the circuit speed. However, by changing $V_T$ we are able to modulate the drive strength allowing for compensation through body biasing. Voltage changes can be a very slow process, for example due to a battery slowly discharging with an application depending rate from years to hours. For energy harvesting applications such as a solar cell, we can expect to see different illumination levels throughout the day, imposing the requirement to follow the changes in the environment. The highest regulation effort can be expected in applications where biasing is intended to correct for fast changing voltages like voltage ripple imposed onto the circuit supply voltage by imperfect regulators.

Third, the temperature has to be taken into consideration: The main effect of temperature is an increase sub-threshold current where temperature is introduced as the $kT$-term in (1.9). Hence, the warmer the circuit gets, the higher the leakage through the the "closed" transistors becomes. However, there is also an effect on the speed, manifesting as a reduction in threshold voltage as well as carrier mobility [43] which can be compensated through the application of body biasing.

For low power circuits the temperature is imposed by the environment and typically changes at a fairly slow pace. Hence, regulation can be implemented with a regularly scheduled calibration with a duty cycle in the range of minutes to hours to months, depending on the application. For high performance circuits we see a different profile: the circuit consumes a significant amount of power, resulting in a self heating effect [44], [45]. Hence, biasing-based compensation would have to follow the power dissipation profile of the application, requiring fast regulation in the range of seconds to microseconds, potentially in a more fine grained manner only for the currently active cores or accelerators in the context of a large SoC.

However, in many cases the requirements for the regulation are actually driven by the application requirements and the power management strategy if the biasing is used to trade off circuit speed against power during operation. In these cases the circuit will switch between different operational modes, such as an ultra low power retention mode, short idle periods, and active high throughput computation with duty cycles dictated by factors such as sampling rates, algorithm complexity, or user interaction. Furthermore the specificity of the circuit to a given purpose plays a role: An ASIC for a well defined single task, always running the same algorithm allows for a highly, at design time, optimised biasing system. Contrary, the designer of a general purpose micro controller needs to expose enough flexibility for the enduser to optimise the operation to their specific application.

## 4.2 Components of a Biasing System

In principle a biasing system consists out of three main components: first, a sensing element for a biasing sensitive target quantity. Second, some kind of control system deciding on changes comparing the sensor data to a set point. Finally we have the bias generators responsible for placing the PWELL and NWELL of the NMOS and PMOS transistors to the appropriate potential.

### 4.2.1 Sensor

Typically the biasing system control is adjusting the circuit speed so that it matches the frequency required for the current task. As shown in Fig. 4.3 the critical path length must be shorter than the period of the clock in order to guarantee correct operation of the digital circuit. On the other hand the critical path lengths should not be significantly shorter than the clock period due to power considerations: the more forward the digital circuit is biased to shorten the critical path the more leaky the circuit becomes.

In order to provide the reference for the bias control the sensor has to measure some quantity relating the length of the critical path with the applied bias. Applicable strategies include Razor-like direct timing error detection [22], approaches based on ring oscillator frequencies, critical path replicas [46], or time to digital converters as well as direct on current sensing

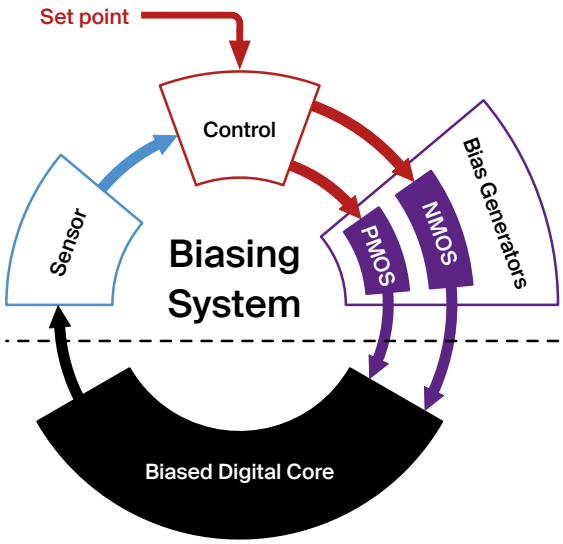



Figure 4.2 – Components of a biasing system for digital circuits.

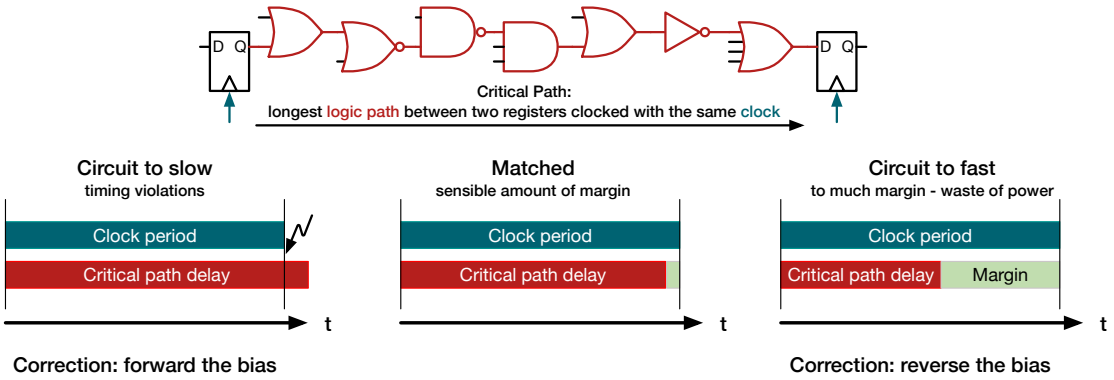



Figure 4.3 – Body bias based critical path matching

through replica transistors.

### 4.2.2 Bias Generators

The Generator is driving the wells towards a bias voltage. The P- and NWELL are imposing a mostly capacitive load to the driver with typically a fairly low leakage which has to be sustained. Hence, the drive requirements for setting the well potential to the body bias voltage are limited for moderate duty cycles. Typically the NMOS bias needs to be able to go negative for reverse operation which comes with the overhead of implementing a charge pump. There are multiple strategies for bias voltage generation described in literature: the potential can be directly supplied through an operational amplifier (for example [47]), or the wells may be shorted towards the rails in a push/pull manner with tightly controlled timing, potentially directly employing a charge pump onto the wells (for example [48]).

### 4.2.3 Control

The control strategy highly depends on the choice for the bias generation: in case of a continuous opamp regulation the control could be as simple as a direct analog feedback loop with the sensor. Charge pump / switching based systems tend to require more complex control mechanism with guaranteed timing. This approach typically requires a digital control system, in particular if precise timing is needed.

An important factor to consider is the need for separate control of PMOS and NMOS: both wells are separately biased, requiring for some method to guarantee a reasonably balanced drive strength for PMOS and NMOS transistors. To complicate things even further the two wells are interacting with each other by construction: The NWELL sits inside the PWELL with a reverse biased PN-junction between the two. This limits the useful bias range to just as much forward where this diode is not conducting. Furthermore this diode is large and thus results in a non negligible capacitance between the two wells. In practice regulation on one well will cause a drift of the other well if left floating, resulting in a shift typically in the opposite direction of the intended bias. Hence both wells have to be regulated jointly.

## 4.3 Implementations and Applications

Several approaches have been discussed in literature with a fairly broad spectrum of applications and goals in mind. This section will break down the different design goals for which biasing systems were considered and reference to example implementations

### 4.3.1 Static Process Compensation

A common application of body control is process compensation. The application of adaptive body biasing has been pioneered by Tschanz et al. [18], showing how a simple, PMOS-only, system can be employed to reduce the spread of both die-to-die as well as intra-die variation to increase the number of dies placed inside their chosen high frequency bin. Abouzeid et al. [49] show a static approach where, based on process outcome, slow chips are pulled back into specification with the application of a static forward bias, resulting in a 30% improvement of static power over a conventional CMOS design in the same technology.

### 4.3.2 Short Term Retention

Next, Biasing can be used to implement retention modes. This has been shown by Blagojević et al. in [48]: A fast body biasing generator using a 1 GHz sampling and charge pump clock is capable of changing between bias modes within a few ns. This allows the circuit to take advantage of short idle periods by rapidly switching in and out of retention modes. A similar approach is proposed by Rossi et al. [50], implementing short low power "power naps" for idle

periods where a state dump to NVM would not be cost efficient while also using the generator to push overachieving fast corners back to reduce leakage. Sensing is done using dedicated process monitoring blocks for N- and PMOS with the loop closed through software.

### 4.3.3 Joint Operation with Supply Voltage Scaling

Further, biasing can be used in combination with other power saving techniques such as supply voltage and frequency scaling to further reduce the consumption. S.M. Martin et al. presented in [51] a concept to reduce power in high-performance processors, resulting in an energy reduction of 48% over the pure voltage scaling approach. In [47] Meijer et al. present a forward body biasing generator, based on two separate DACs for NMOS and PMOS respectively. Limiting the generator to forward only simplifies the generator design significantly as there is no need to generate negative bias voltages for the NMOS. Their approach allows to shift the circuit in the $V_{DD}$-$V_{BB}$ supply voltage design space based on a LUT approach. Recently, Lee et al. [41], [46] presented a scheme controlling the bias jointly with the supply voltage based on the ratio between leakage and dynamic power in DDC. They showed that this ratio closely tracks the minimum energy point across PVT, allowing for its use as an easier to measure proxy. Their novel proposal utilises a counter tracking the duty cycle of the DC/DC converter at two different core frequencies in order to determine the leakage to dynamic ratio. Finally, show that their system is capable of tracking the MEP by dynamically adjusting the bias and supply voltage when sweeping the temperature.

### 4.3.4 Design Time Optimisation with Multiple Bias Partitions

Finally, biasing has been proposed as a tool for system power optimisation at design time. The idea behind this approach is to break down a complex circuit into partitions of different bias voltage in order to trade of speed against leakage power. This approach has been followed by Kulkarni et al. [52], using a greedy clustering algorithm placing strongly connected gates in separately biased partitions in order to allow for intra die compensation through dynamic biasing. Similarly, Kühn et al. [53] propose a bias domain partitioning scheme in order to explore the potential of design time partitioning into bias domains based on the idea of dissimilar length of critical paths in consecutive pipeline stages.

## 4.4 Biasing Concepts in this Thesis

In the following we will explore biasing concepts for the three chips presented in Tab. 4.1 with different biasing implementations:

Calanda, is a general purpose ultra low power micro controller. Its biasing system uses the on-current as a proxy for the circuit speed and allows the user to control it through a current DAC, providing a reference fed into a comparator against the currents observed through diode

Table 4.1 – Comparison of the biasing system implementations for DDC

| | Calanda | Nakayama | Sneafellsjokull |
|---|---|---|---|
| Bias System | $I_{ON}$ current control | $I_{ON}$ current control & FLL | Charge pump based |
| Sensing | $I_{ON}$ monitors | $I_{ON}$ monitors & reference clock counter | NMOS/PMOS balance sensor |
| Set point | IDAC value | Frequency | Predefined operation and retention mode |
| Frequency | Fixed & supplied externally | Generated on chip and regulated for current bias | Two fixed, one for retention, one for operation |
| Supplies | Fixed | Adjustable | Fixed |
| Application | General purpose MCU | General purpose MCU with full flexibility in operation | Special purpose MCU with well chosen operating point. |

connected transistors of the same sizing to the ones used in the standard cell library.

Nakayama extends the concept from Calanda by adding a secondary control loop on top of the current regulation: The user directly sets a target frequency for a frequency locked loop with a standard cell based oscillator. The oscillator frequency is referenced against an external 32 kHz with a counter acting as the sensing element for the bias control loop. The FLL then directly controls the DAC codes, adjusting the bias accordingly to achieve the requested frequency.

Finally, Snaefellsjokull implements a simplified biasing scheme using a switch to select between two predetermined PMOS bias supplies together with a charge pump controlled by a PMOS/NMOS balance sensor. We show how this simple concept, together with a well chosen operating point, is capable to reduce the worst case power significantly while achieving a reasonably low penalty over a more complex system.

All three Chips have been manufactured and the results presented are, unless stated otherwise, measured for both Calanda and Nakayama. Unfortunately the values shown for Snaefellsjokull are all based on simulations due to a short between two wells in the layout which was missed during the verification rendering the NMOS bias charge pump nonfunctional.

# 5 Calanda: Analog On-Current Regulation

Low power micro-controllers typically [4], [13], [34], [54]–[56] employ supply voltage scaling for compensation of process variation as well as for trading off power against performance. This can be a very efficient method, however, it comes with the need of highly adjustable voltage converters, capable to implement the scaling over a wide range while also staying efficient both during low current retention and in high performance modes when high performance is required.

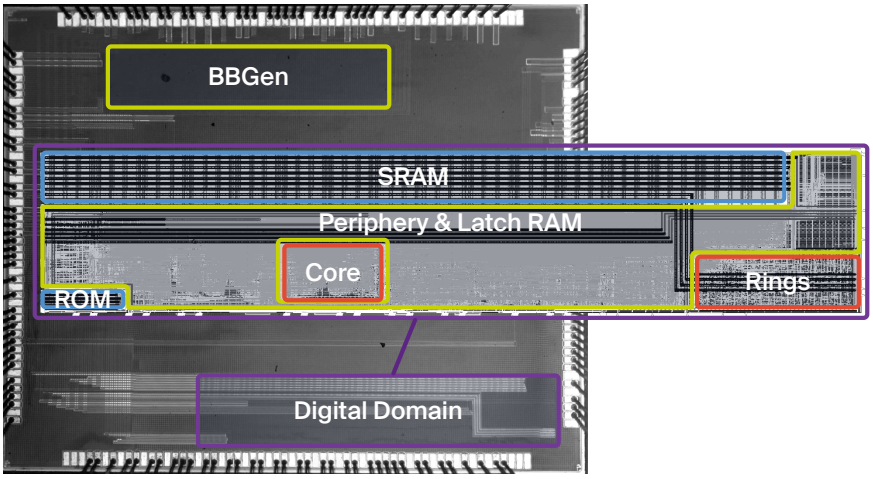



Figure 5.1 – Annotated die shot of the Calanda SoC.

Calanda, shown in Fig. 5.1, was designed with the goal to avoid supply voltage scaling to provide a wide range of energy-optimal frequency operating points by exploiting the large body factor of 55 nm DDC . This concept allows for the use of one, highly optimised, single voltage DCDC converter to supply the SoC, while the biasing circuitry only sees a small, mainly capacitive, load allowing for a compact implementation.

A common method of constructing biasing generators is based on a DAC, setting the NMOS and PMOS bias voltages directly [46]–[48], [50]. However, with this approach variation is not

implicitly considered and the loop is commonly closed through a digital control system fed with the output of some kind of timing monitor. Most timing based approaches, such as ring oscillators or critical path replicas, only sense the aggregation of NMOS and PMOS variation which will result in non-optimal compensation for imbalanced process corners (SF/FS).

Instead, Calanda uses a novel concept developed within CSEM [57], that utilizes a current DAC to provide a reference current which is then matched by applying bias into pull-up and pull-down circuit replicas for PMOS and NMOS respectively.

Within this chapter we first explore the system architecture, and describe the biasing concept. We then describe a ring oscillator based test infrastructure integrated on chip.

The second part of the chapter explores the measurements of the chip. We describe the measurement setup and explore the biasing design space. We then show measurements of the ring oscillators and the core, verifying the biasing system. We then use the ring oscillators to verify the ADVBB model presented in Chapter 3, showing a well matched prediction of a constant frequency trajectory as well as a near perfect match for dynamic power and an acceptable match for leakage. We close with system results, exploring the operable SRAM bias design space and providing core power measurements.

Parts of this chapter have been published in the following paper:
M. Pons, C. T. Müller, D. Ruffieux, *et al.*, "A 0.5 V 2.5 $\mu$W/MHz Microcontroller with Analog-Assisted Adaptive Body Bias PVT Compensation with 3.13nW/kB SRAM Retention in 55nm Deeply-Depleted Channel CMOS", in *2019 IEEE Custom Integrated Circuits Conference (CICC)*, Apr. 2019, pp. 1–4. DOI: 10.1109/CICC.2019.8780199

## 5.1 System Architecture

The Calanda system, shown in Fig. 5.2 is an SoC integrated in 55 nm DDC, intended as a test bed for an on-current based biasing system.

The system is built around a 32 bit RISC Icyflex core [58]. For full operation 64 kB of SRAM are available to the programmer while an additional 4 kB of latch based standard cell memory placed in the same bias domain as the core allows for limited operation in retention without the need for SRAM. Finally there is a 4 kB boot ROM available, implementing a boot sequence from an external SPI flash. Finally JTAG allows for on chip debugging and standard periphery such as UART, SPI, Timers, GPIO is also integrated.

Separate Biasing systems are integrated for the core, the SRAM bitcell and periphery, the ROM as well as a test structure domain containing a collection of standard cell based ring oscillators.

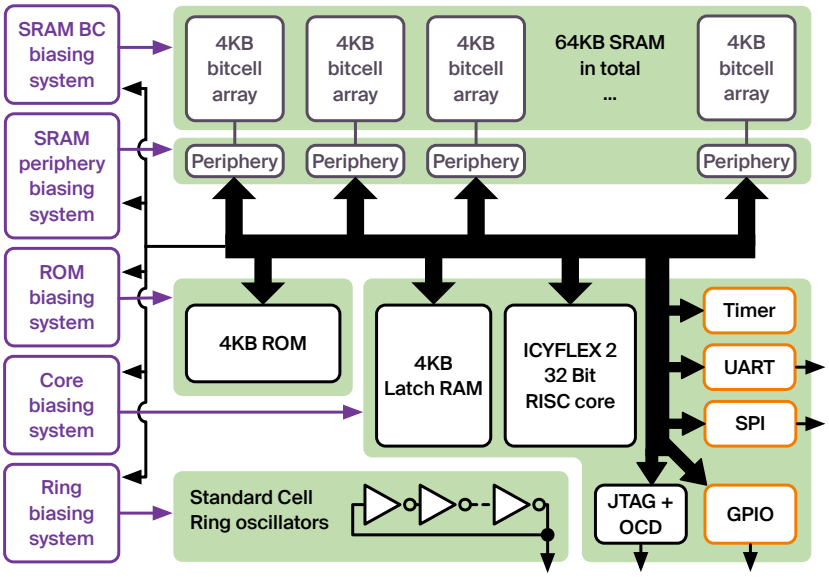



Figure 5.2 – Calanda System overview: The system integrates a 32 bit RISC Icyflex 2 core together with 64 kB of SRAM, 4 kB of Latch SCM, 4 kB of ROM together with standard periphery such as UART, SPI, Timers, GPIO as well as JTAG. Sepparate $I_{ON}$-Current regulating biasing systems are available for the SRAM, CORE, ROM, and a collection of standard cell based ring oscillators.

## 5.2 On-Current based Biasing Concept

The concept of Calanda is based on the idea that, as a first order approximation, a large share of process, temperature and supply voltage variation manifests in a change of the on-current $I_{ON}$ that can be driven by the transistor. Hence, by designing a biasing system that controls $I_{ON}$ to a set point across PVT we can expect to cover a large share of the variation and achieve a fairly close match in standard cell timing.

In essence, Calanda replaces the commonly used circuit delay reference sensor with a current sensor. The sensor output is compared to a current reference that determines the Ion set point. This current based approach is easy to implement for both PMOS and NMOS, which avoids the issue of timing based sensors that can not easily separate the delay arising from either of the two. This is crucial in digital CMOS circuits as, by construction, we observe an alternation between pull-up and pull-down networks when traversing a critical path. Balancing is of particular importance for FS or SF process corners where a significant miss balance is built in [59], or for circuits operated in the sub-threshold regime where even a small amount of imbalance can have a significant impact [60].

The heart of the biasing system of Calanda is shown in Fig. 5.3 (compare also [57]). The control is achieved by programming a reference current digitally via two 8-bit current DACs, which provide the reference for NMOS and PMOS respectively. The current imposed by the DACs is then fed into monitors which form the sensors of the bias system. The monitors are implemented as simple pull-down and pull-up replicas, which act as sensors for the bias domain and are

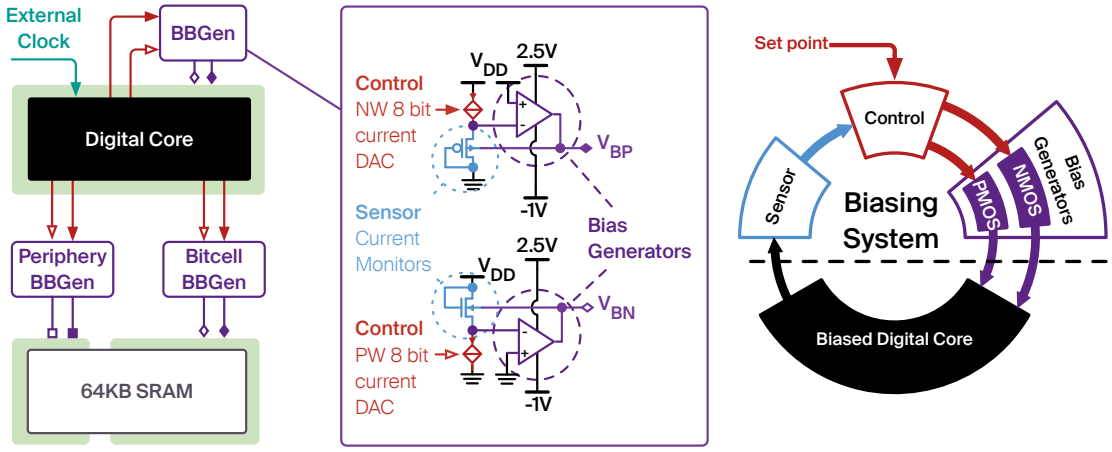



Figure 5.3 – Calanda Biasing System: An operational amplifier controls the bias such that the on-current through representative on-current monitors matches the one set by a current-DAC.

implemented using diode-connected PMOS and NMOS respectively. The connection is such that the replicas share the same $V_{GS}$ and $V_{BS}$ as the actual circuit biased, guaranteeing the same bias condition. The replicas are sized to share the same $W/L$-ratio of $^{205\,nm}/_{90\,nm}$ used as a base unit in the standard cell library. The long channel has been chosen to reduce worst case leakage. Finally, an operational amplifier is used as the bias generator. Its task is to adjust the bias voltage just so that the current through the replica matches the current provided by the current-DAC. The output is the bias voltage which is fed into the pull-up/pull-down replicas as well as into the to be biased power domain.

In order to achieve compensation the current reference has to be designed to be variation invariant. This is a common problem with existing solutions readily available in the analog designers toolbox [61], [62, p. 315ff.]. Once a PVT invariant reference current has been established it can be copied with current mirrors, allowing to trivially extend the reference into a DAC. Common approaches use parallel unit currents which are switched in either a binary weighted or unary fashion before being summed up [63, p. 114ff].

However, the approach based purely on $I_{ON}$ has, the disadvantage that the effect of $V_{DD}$ on (1.4) from the introduction is not considered. This approximation in an undercompensation for too high supply voltages and an overcompensation for too low supply voltages due to the $^1/_{V_{DD}}$ relationship.

$$f \propto = \frac{\overbrace{I_{ON}}^{\text{PVT compensated}}}{C_L} \cdot \frac{1}{\underbrace{V_{DD}}_{\text{uncompensated}}} \tag{5.1}$$

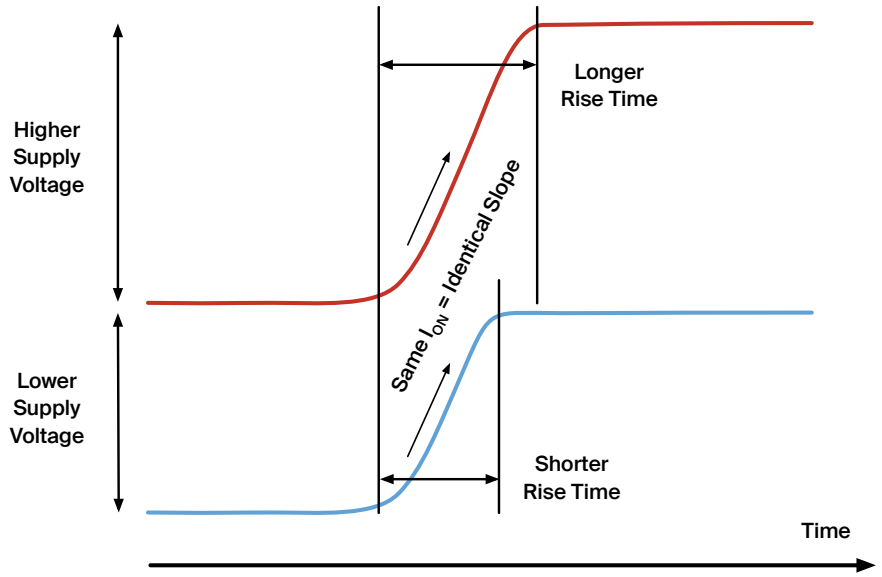



Figure 5.4 – Sketch of the rise time for constant $I_{ON}$ at different supply voltages: as $I_{ON}$ sets the slope rather than the transition time a higher supply voltage results in a slower circuit.

Intuitively this can be understood by considering that $I_{ON}$ modulates only the slope of the rise and fall transition as sketched in Fig. 5.4. For identical $I_{ON}$ we see an identical slope and hence transitions for a higher supply voltage take longer, resulting in a slower circuit speed.

Indeed we have seen already in Chapter 2 based on standard cell characterisation data that the assumption that we can compensate by forcing a given $I_{ON}$ holds: the compensated corners shown in Fig. 2.4 were in fact biased in such a way that the $I_{ON}$ for NMOS and PMOS across the corners was equalised. Also, we can observe that the remaining variation is roughly following $1/V_{DD}$, resulting in a slight overcompensation equivalent to the 10 % variation in voltage. However, this effect is small enough compared to the effect of $V_{DD}$ on $I_{ON}$ in the near- and sub-threshold domain that it can be neglected for the sake of a reduced implementation complexity by just considering an MMMC approach during implementation of the digital design in order to provide a sufficient amount of margin.

Let us now have a look into some power considerations. In particular, we consider not only leakage and active power, but include also the power required for driving the well to apply the desired body bias voltage. For forward operation power is dominated by the dynamic component, with a consumption far higher than the static currents expected driving the bias voltages. However, when implementing a retention mode, the large bias range is only useful as long as the power contribution of the biasing system in reverse operation is lower than the corresponding power savings in the biased domain. This might be a concern due to the relatively large size of the reverse diodes forming the wells.

Figure 5.5 shows the circuit leakage in relation to the well leakage for an SRAM bit cell. As we can see, the well currents reside well below the leakage currents of the cell, with the only

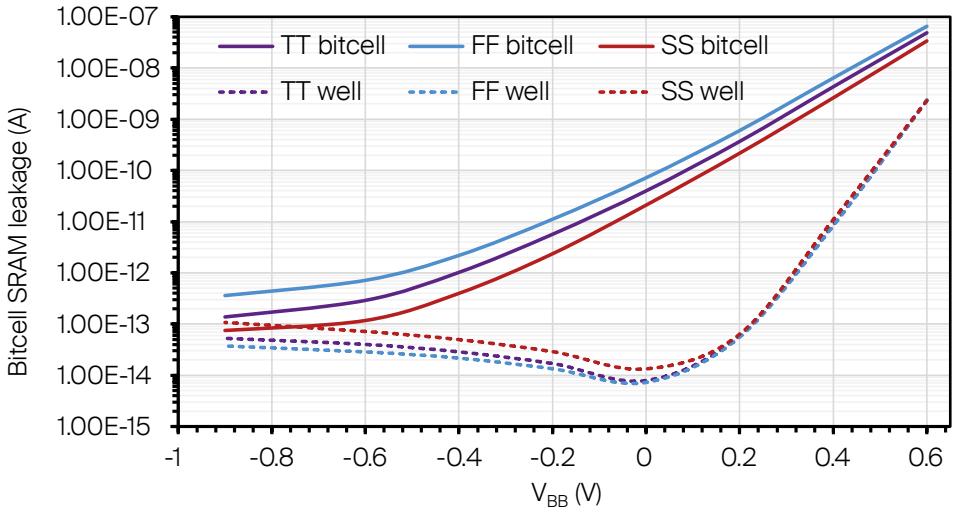



Figure 5.5 – Well currents in relation to the leakage currents for an SRAM bit cell (data provided by USJC)

exception of a slow chip in a far reverse condition. That configuration would typically not be reached in an adaptive system as the slow chip would be forward biased to compensate. Hence, we can assume that power consumption the bias generation will not dominate retention power.

## 5.3 Standard Cell Ring Oscillators for Bias System Characterisation

As we have seen in Chapter 2, we can expect some variation in the scaling across PVT conditions for different standard cells. This is mainly due to the different internal structure of the cells, manifesting as differences in stack heights and sizing for their particular logic function. We have also seen that the median over all cells tracks the overall circuit speed reasonably well. Hence, we are interested in a test structure based on a reasonable selection of different standard cells to quantify the remaining variation of the system after applying the $I_{ON}$ based biasing approach.

As it is not feasible to implement test structures for each cell in the library, measuring timing for each possible input transition and output load a simplified approach, has to be considered: We can assume that the most commonly used cells are be responsible for the largest share of variation seen throughout a typical design. Sequential cells such as flip-flops and latches can be omitted due to the expected small impact on the total delay of long paths which simplifies the task as specialised test circuitry would be needed to extract setup and hold constraints.

Furthermore, when constructing test structures the fan out has to be taken into consideration. Synthesis and place and route tools will typically select large buffers for driving high fan out

nets and global routing. Contrary, low drive strength cells are used for local logic clusters where predominantly low fan out nets are found. Hence it is necessary to load the cells similarly to the conditions they would be used in within the actual circuit.

Ring oscillators can be used as a simple model for delay which is also straightforward to implement in hardware and hence have been used frequently for tasks like variation analysis [64], on-chip performance monitoring [65] or as a base for system modelling [66] in the context of body biased systems. Their delay can be easily evaluated by just measuring the frequency. For Calanda this measurement has been done externally, but if a use as an on chip timing sensor is desired there are options for full integration. The most simple option uses a counter, relative to a known external reference clock, and can provide very accurate results when averaged over a reasonably large number of cycles. The drawback of this technique over the second option, using a dedicated time to digital converter, is the relatively long cycle time, which does not allow to react on fast events such as voltage droops. These fast events however are typically not an issue for the slow circuits in near and sub-threshold regime considered in this thesis.

### 5.3.1 Ring Oscillator Construction

In order to select a representative set of cells the distribution of cells has been extracted in an early design stage from the post synthesis netlist of the Calanda system. The top 25 combinational cells used in the netlist have been selected to construct ring oscillators. Table 5.1 lists these cells in order of their number of instances within the test design, showing that these 25 cells cover slightly more than 94% of the cells used in the actual design. Each instance sees a slightly different input slew and output load.

To achieve a realistic load configuration the cumulative load distribution has been derived by adding up the load of each instance in the netlist. As shown in Fig. 5.6 for the two extremes of a drive strength one inverter and a drive strength six buffer we arbitrarily chose the value at 20% of the CDF as a typicall “low” load and the 80% value as a “high” load for that particular cell. As the synthesis tool selects the cells drive strength appropriately for a given load, this approach guaranties that cells are analysed in the load condition for which the cell was designed.

Based on the logic function of the cell a suitable arc has been selected arbitrarily to either implement an inversion or a propagation of the input. One hundred instances of each cell have then been chained up and an additional NAND gate was added in the end of each chain to implement the inverted feedback as well as an enable to silence unused chains. Finally, each of the hundred cells has been loaded with dummy cells to achieve the 20% and 80% loads derived from the CDFs. “Don’t touch” attributes have been used in order to ensure that optimisation throughout place and route does not remove these load cells.

The standard cell oscillators have been placed automatically in their own power domain with the intention that automated placement over the 100 cells will result in an average routing resembling something similar seen in a digital domain.

Table 5.1 – Combinatoric cells used for the construction of ring oscillators with their 20% and 80% loads derived from the cumulative distribution function.

| Cell | Drive Strength | Instances | Percentage | $C_{CDF20}$ (pF) | $C_{CDF80}$ (pF) |
|---|---|---|---|---|---|
| INV | 1 | 44488 | 17.75 | 0.001103 | 0.003688 |
| BUF | 6 | 36532 | 14.57 | 0.071947 | 0.122444 |
| NOR2 | 2 | 29960 | 11.95 | 0.001103 | 0.002394 |
| NOR2 | 3 | 19071 | 7.61 | 0.001886 | 0.003056 |
| NAND2 | 2 | 17372 | 6.93 | 0.001046 | 0.003497 |
| AND2 | 1 | 11775 | 4.70 | 0.002952 | 3.346537* |
| NAND2 | 3 | 11081 | 4.42 | 0.001857 | 0.004355 |
| INV | 3 | 10939 | 4.36 | 0.004280 | 0.010376 |
| INV | 2 | 6170 | 2.46 | 0.002496 | 0.011044 |
| AN22 | 1 | 5537 | 2.21 | 0.000736 | 0.001119 |
| BUF | 4 | 5446 | 2.17 | 0.006657 | 0.023673 |
| ON22 | 3 | 4476 | 1.79 | 0.001857 | 0.002394 |
| AND2 | 2 | 3675 | 1.47 | 0.011400 | 0.023490 |
| OR2 | 1 | 3599 | 1.44 | 0.000985 | 0.006648 |
| AND2 | 3 | 3268 | 1.30 | 0.011779 | 0.028453 |
| ON22 | 2 | 3128 | 1.25 | 0.001857 | 0.002394 |
| MUXI2 | 1 | 3045 | 1.21 | 0.000793 | 0.001534 |
| NAND3 | 1 | 2917 | 1.16 | 0.000737 | 0.002081 |
| AN21 | 1 | 2851 | 1.14 | 0.000738 | 0.001422 |
| INV | 6 | 2469 | 0.98 | 0.285539 | 0.309147 |
| ON21 | 1 | 2274 | 0.91 | 0.000742 | 0.001877 |
| NAND3 | 3 | 1781 | 0.71 | 0.001857 | 0.002394 |
| NOR3 | 1 | 1490 | 0.59 | 0.000736 | 0.001598 |
| OR2 | 3 | 1286 | 0.51 | 0.011001 | 0.030628 |
| XNOR2 | 1 | 1204 | 0.48 | 0.000742 | 0.002249 |
| covered | | 235834 | 94.07 | | |
| remaining | | 14839 | 5.93 | | |
| total | | 250673 | 100 | | |

*The value for the AND2 is heavily skewed by its use as a clock gate on the post synthesis netlist, the actual hardware was designed using a CDF80 value of 0.00781, based on the value of an other cell with a similar CDF20 value.

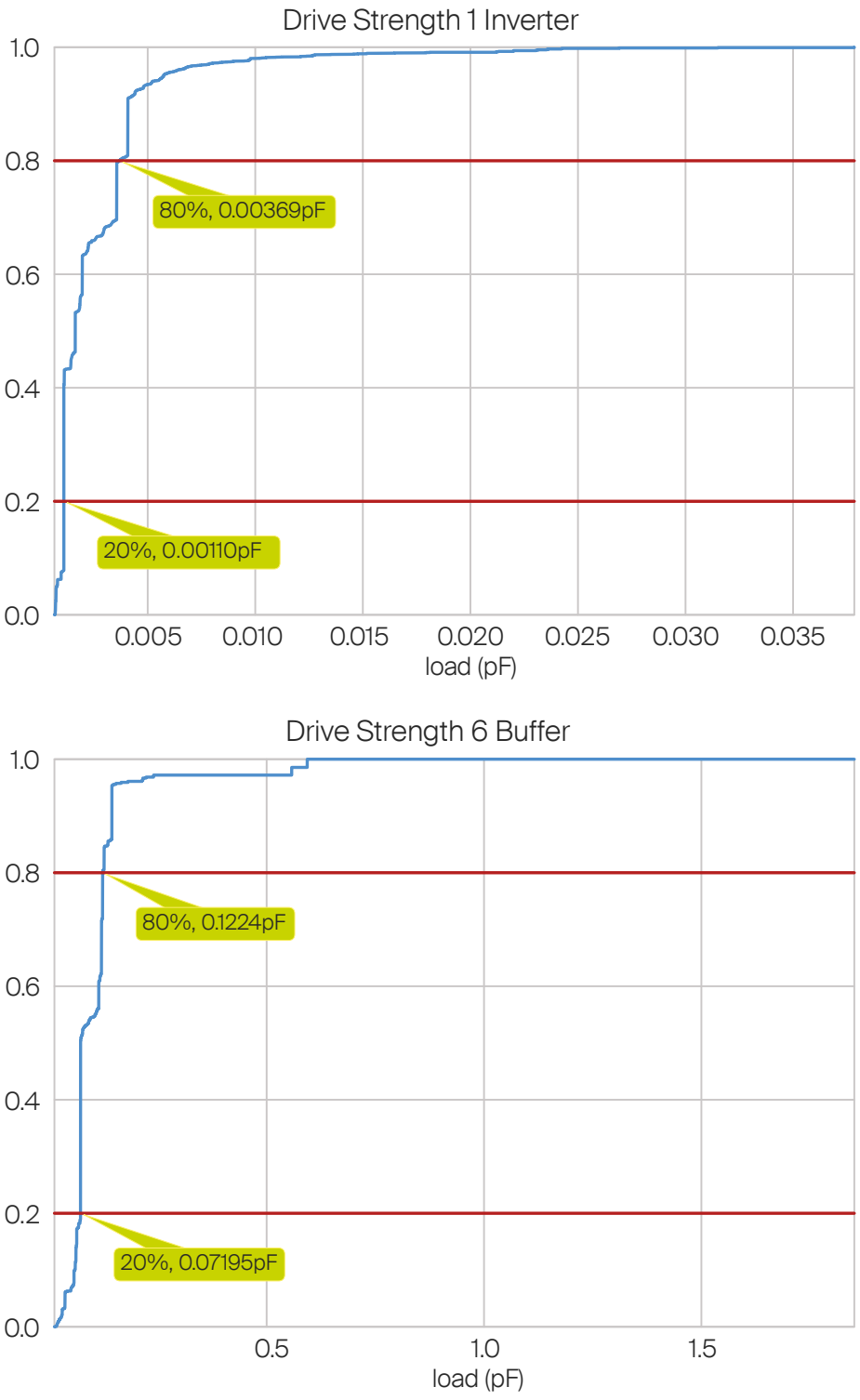



Figure 5.6 – The cummulative distribution functions of the load for the two most commonly used cells of the used test design. The stronger drive strength 6 buffer is loaded with a much higher load than the small drive strength one inverter.

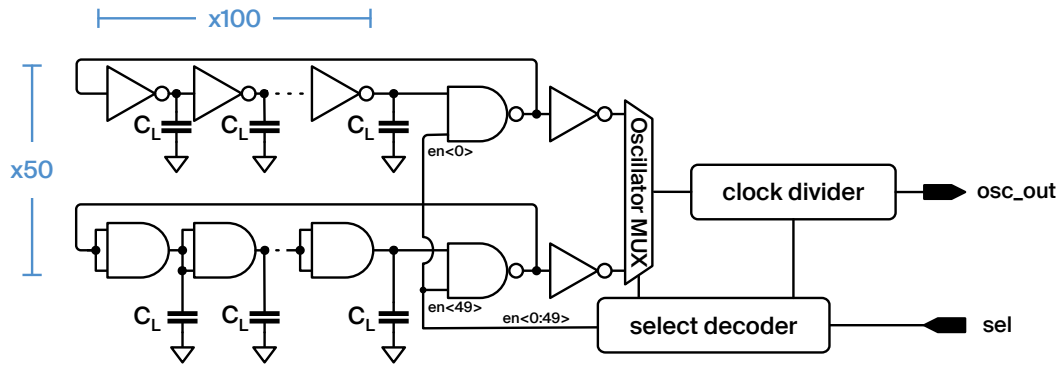



Figure 5.7 – Standard cell ring oscillator test architecture: 50 rings of the length of standard cells. The rings are based on 25 different cells, loaded with a typical high and low load for each particular cell.

As shown in Fig. 5.7, each oscillator is silenced when not active using a final NAND gate. A MUX selects the output of the currently running oscillator and the optionally frequency divided output is placed on a pad for analysis with a frequency counter.

## 5.4 Chip Measurements

In this section we present measurements of the Calanda SoC. We will quickly introduce the measurement setup used, follow on with an exploration of the design space of the $I_{ON}$ currents fully reachable across PVT, setting the boundary for the compensation. We follow up with a closed loop analysis, showing the behaviour of the rings and the core under control of the Calanda biasing system. We then verify the model from Chapter 3 in an open loop fashion, using the model as an oracle for the correct bias values for a constant frequency trajectory through the body bias supply voltage design space. We close with a system perspective, presenting both core and SRAM results across PVT for two different operating modes.

### 5.4.1 Measurement Setup

All following measurements have been done with the setup depicted in Fig. 5.8. An Agilent B2902A SMU has been used to supply and measure the core power. For open loop measurements a pair of Keithly 2401 SMUs is used to provide bias voltages with precise current measurements while two PXI-4130 SMU cards were used to provide the bias to currently not measured bias nets. As the system features many bias domains we used a PXI-2530 switchbox with an 8x16 matrix terminal block to allow to connect and disconnect the the nets to the SMU of choice for open and closed loop operation respectively. In addition a Keithly 2000 multimeter is connected to the switchbox, used during closed loop measurements to characterize the bias voltages. A second switchbox is used with an additional Keithly 2400 SMU for characterizing replicas of the current monitors. The System clock is generated externally with an Agilent

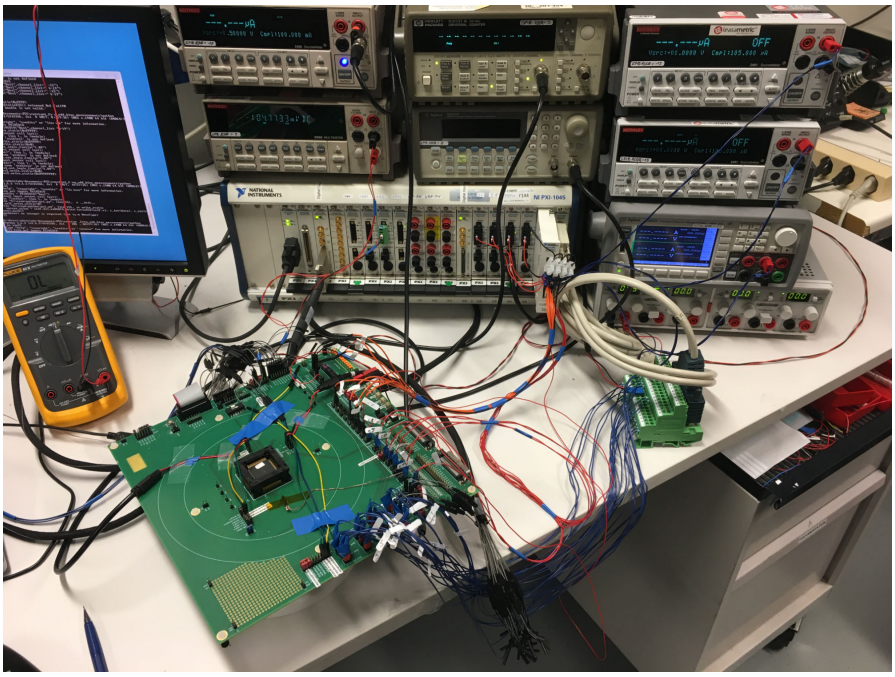


Figure 5.8 – Measurement setup for Calanda.

33250A arbitrary waveform generator while the ring frequencies were characterised using an HP 53131A frequency counter. GPIOs were controlled through a PXI-6552 digital IO card, which was used for implementing handshake protocols between software running on the core and the higher level measurement scripts. The system was controlled through an Amontec JTAGkey. Finally, a Temptronic TP04310A thermo cycler was used to cool and heat the device under test (DUT) to −40 °C and 85 °C, respectively while measuring the temperature through a sensor which was placed on the package.

All GPIB based instruments were controlled through the PyVISA API [67] while the PXI devices where controlled through the DLL drivers provided by National Instruments. OpenOCD [68], [69] was interfaced through the RPC interface, implementing direct read/write of the system memory bus of the SoC. This proved to be very powerful and flexible, as it allowed to reconfigure the SoC through direct access to the configuration registers.

The chips were manufactured by USJC as split lots which allows to measure the design in SS, TT, and FF conditions.

### 5.4.2 Biasing Design Space

In Chapter 3 we have shown that USJC 55 nm DDC leaves a wide range for the designer to trade off frequency against power, however that analysis was done purely on model data. In the following we first verify the predicted achievable $I_{ON}$-scaling observed on silicon.

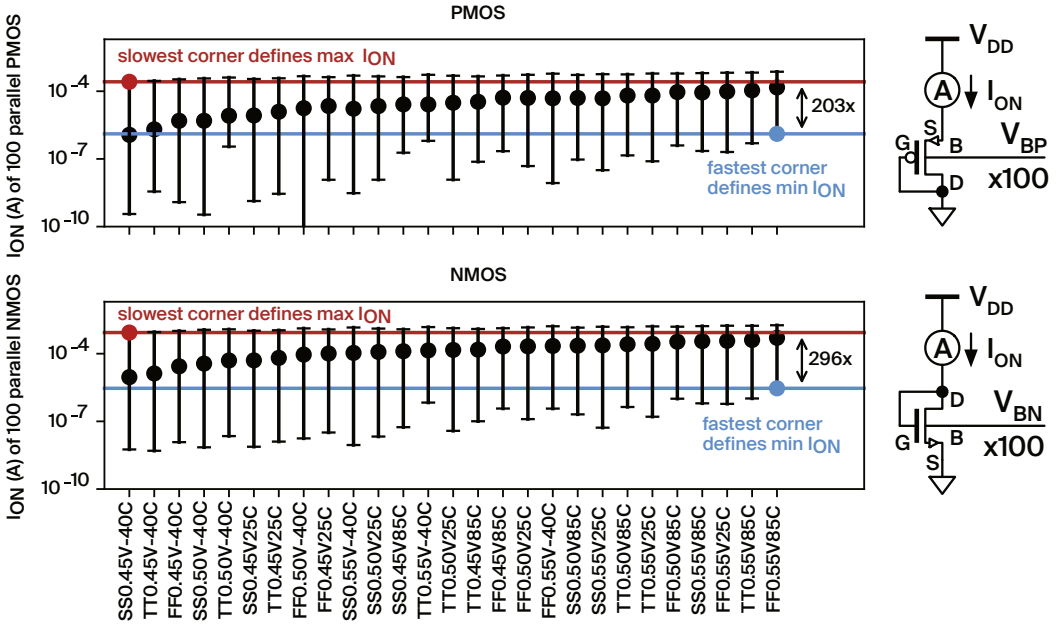



Figure 5.9 – Measured $I_{ON}$ currents through 100 parallel W/L=205/90nm N and PMOS ULL DDC transistors: TT, SS, FF samples, $V_{GS}$=$V_{DS}$=0.5 V±10%, −40 °C, 25 °C, 85 °C. Considering PVT compensation, $V_T$ tuning capability by body bias allows $I_{ON}$ scaling of ≈ 200× for PMOS and ≈ 300× for NMOS.

To this end a test structure was implemented, integrating 100 parallel diode connected NMOS and PMOS ULL transistors. The transistors were integrated identically to the ones used as current monitors in the biasing system, i.e. with a width to length ratio of 205 nm to 90 nm which the same value used as minimal dimension in the standard cell library used for Calanda. The bias and supply voltages were supplied externally, the latter one with a Keithly 2400 series SMU, allowing to measure the on-current directly. In particular three different operating points were targeted: fully reverse, no bias, and fully forward.

The results are shown in Fig. 5.9, depicting the measured $I_{ON}$ across three process corner lots (SS,FF,TT), with the supply voltage varying by 10 % around the nominal 500 mV, and at temperatures of −40 °C, 25 °C, and 85 °C. The upper and lower markers here represent the highest and lowest $I_{ON}$ achievable respectively while the black dot marks the no bias condition i.e. with PMOS at $V_{DD}$ and NMOS at $V_{SS}$.

The design space spanned by $V_T$ tuning through body biasing for full compensation is now defined by the worst cases: The upper limit is set by the corner with the lowest current that achieved while pushing the circuit into a full forward condition (marked in red). Similarly, the lower limit is set by the corner with the highest current in full reverse biasing (marked in blue).

We can see that, even with the requirement of full compensation which drastically reduces the

range for some corners, we still realise a factor of 203x for PMOS and 296x for NMOS allowing for significant throughput scaling *after* compensation. Note that identical sizing is used for NMOS and PMOS for the standard cell library used in Calanda, relying on the biasing to also compensate for the mobility difference. This decision means that the bias ranges depicted in Fig. 5.9 for NMOS and PMOS have to be overlapped, only leaving the range that is reachable by both transistors. Effectively this results in a further reduced range. The remaining factor of 89.66x however still allows for plenty of speed vs. energy tradeoff.

### 5.4.3 Ring Frequency Across PVT with the Calanda Biasing System

In this section we use the biasing system from Calanda and observe the effect on the rings. We obtain the spread in delay variation of the cells across the corners by first normalising them for their inherent frequency difference due to the use of different cells and load conditions and then with the on-current.

First, each measured frequency $f_r(x)$ for the ring $r$ at the DAC code $x \in X$, where $X$ represents the set of all valid DAC codes, is normalised with the frequency $f_r(\tilde{x})$ at the center DAC code $\tilde{x}$:

$$f'_r(x) \quad = \quad \frac{f_r(x)}{f_r(\tilde{x})} \tag{5.2}$$

This step corrects for the inherrent delay difference between the ring oscillators. In the next step we apply the same normalisation for the currents measured through the monitors used as current sensors:

$$I'(x) \quad = \quad \frac{I(x)}{I(\tilde{x})} \tag{5.3}$$

$I'(x)$ now reflects the current at the DAC code $x$ relative to the current at the center DAC code $\tilde{x}$. The same way, $f'_r(x)$ describes the frequency at the DAC code $x$ relative to the frequency at the center DAC code $\tilde{x}$, scaling the frequency into the same range as the current. We now relate the center-normalized frequencies to the center-normalized currents:

$$r_r(x) \quad = \quad \frac{f'_r(x)}{I'(x)} \tag{5.4}$$

With the assumption from (1.4) we expect r to be a perfect horizontal line at one. Any deviation

indicates a cell-specific deviation from the ideal common delay scaling proportional to the on current through the sensor: a value at 1.1 corresponds to a frequency 10% faster than the expected frequency with $f \propto I$ while a value of 0.9 indicates a 10% slower frequency.

Besides the difference in the delay scaling behavior for different cells for different bias conditions, we are interested in comparing the celse comprising the rings across process $P$, supply voltage $V$, and temperature $T$, defined as the operating corner tuple $\text{þ} = (P, V, T)$. Hence, we extend this normalisation to normalise the rings in other operating corners relative to the typical corner tuple $\text{þ}_{typ} = (\text{TT}, 0.5\,\text{V}, 25\,^{\circ}\text{C})$:

$$f'_r(x, \text{þ}) = \frac{f_r(x, \text{þ})}{f_r(\tilde{x}, \text{þ}_{typ})} \tag{5.5}$$

$$I'(x, \text{þ}) = \frac{I(x, \text{þ})}{I(\tilde{x}, \text{þ}_{typ})} \tag{5.6}$$

$$r_r(x, \text{þ}) = \frac{f'_r(x, \text{þ})}{I'(x, \text{þ}_{typ})} \tag{5.7}$$

### Ring Oscillator Matching

In the following we analyse a data set, generated with a sweep of the DAC codes in three process corners (SS/FF/TT), across the temperature range of −40 °C, 25 °C, and 85 °C as well as over VDD in the range of $\pm 10\%$ around $V_{DD}$. The ring frequencies were directly measured using an external frequency counter with the MUX controlled via JTAG. In addition the corresponding monitor currents were measured across the DAC codes.

Figure 5.10 shows the matching between the rings across the process corners after the normalisation steps using the on-current based compensation approach. The X-axis marks a of of the DAC codes from the bias generator. Everything is normalised to the typical case with a DAC code of 136 near the center.

We observe a fairly close match of the corners, keeping the variation within a band of roughly 10% after compensation except for the far reverse case where the spreads opens up to 40%. This behaviour is expected as the reverse bias forces the operating point of the transistors into the sub threshold domain with an exponential slope. Hence, even small variations in supply voltage or threshold voltage have a significant effect on the timing. A single cell with a bit of local variation can significantly slow down the oscillator. Furthermore, the logic function of the standard cell may play a role: more complex cells contain more transistors in higher stacks. Transistors in stacks do not see a common $V_{GS}$ and the weakest transistor limits the current, resulting in a higher sensitivity to local variations.

Figure 5.11 shows similar results across temperature with a particularly wide spread for the

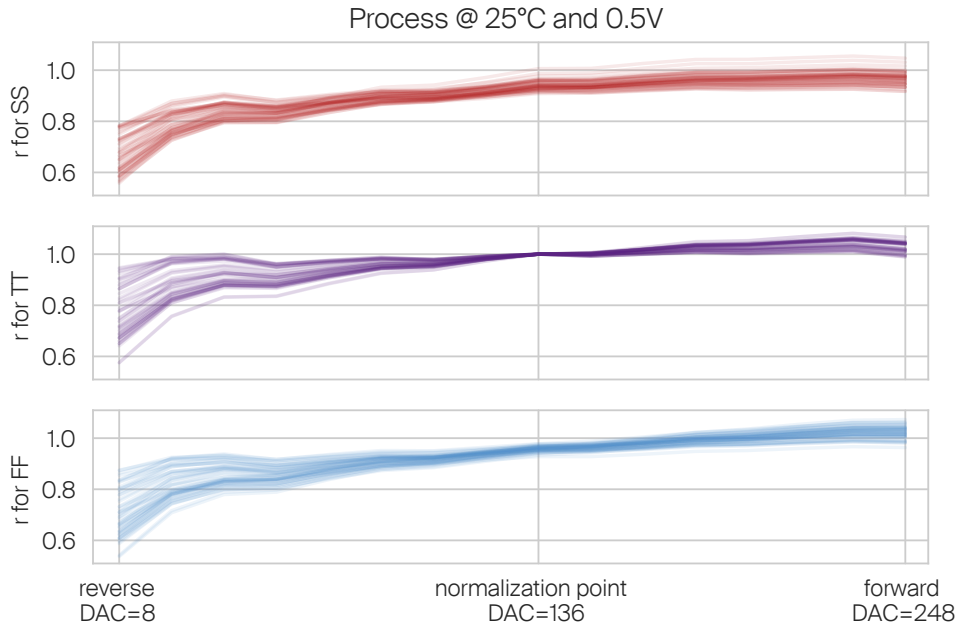



Figure 5.10 – Normalized relative frequency variation of the ring oscillators across process corners when sweeping the DAC code. The order is SS (top/red), TT (center/purple), and FF (bottom/blue).

cell-delay scaling behaviour is observed for −40 °C with reverse bias where a the deviations of individual cells ranges from 60% to nearly 120%.

Finally Fig. 5.12 presents the supply voltage dependency: the general trend follows the observations from the process and temperature quite closely, however we note a distinct offset of the normalized oscillator frequencies from one for all cells which is opposite to the relative offset of the supply voltage. Specifically, at 450 mV, all ring oscillators are approximately 10% faster than at 500 mV while at 500 mV they are about 10% slower than at 500 mV. This observation confirms the overcompensation tendency of the on-current based biasing approach as explained before in Fig. 5.4.

### Core-Ring Variation Matching

In order to achieve a similar characterization for the microcontroller core of Calanda as for the rings, we can adapt Equation 5.5 ff accordingly, with $f_c(x, \text{þ})$ describing the maximum core frequency at DAC code $x$ and at the operating corner þ:

$$f'_c(x, \text{þ}) = \frac{f_c(x, \text{þ})}{f_c(\tilde{x}, \text{þ}_{typ})} \tag{5.8}$$

$$r_c(x, \text{þ}) = \frac{f'_c(x, \text{þ})}{I'(x, \text{þ}_{typ})} \tag{5.9}$$

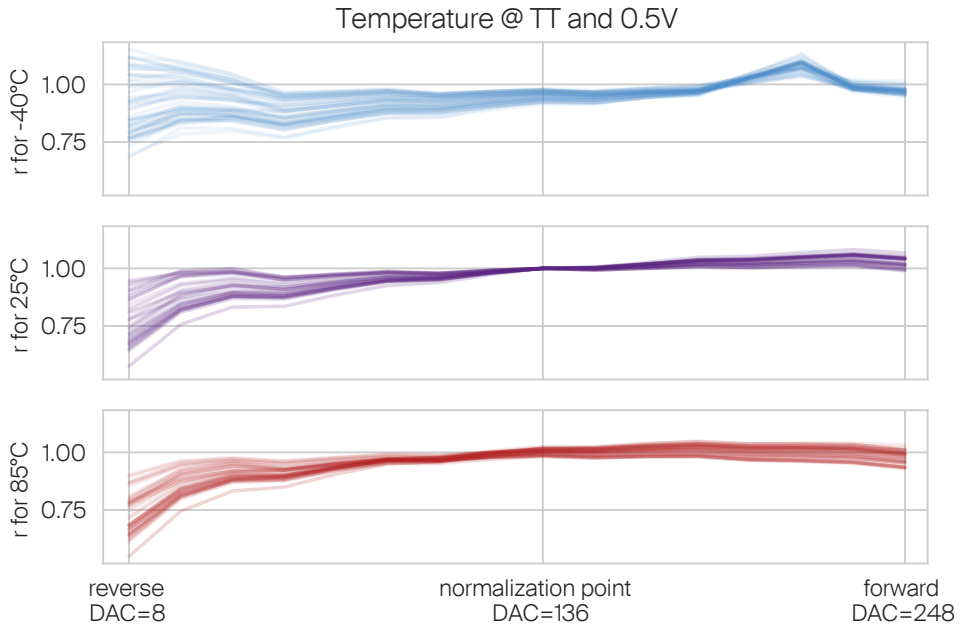



Figure 5.11 – Normalized relative frequency variation of the ring oscillators across temperature when sweeping the DAC code. The order is −40 °C (top/blue), 25 °C (center/purple), and 85 °C (bottom/red).

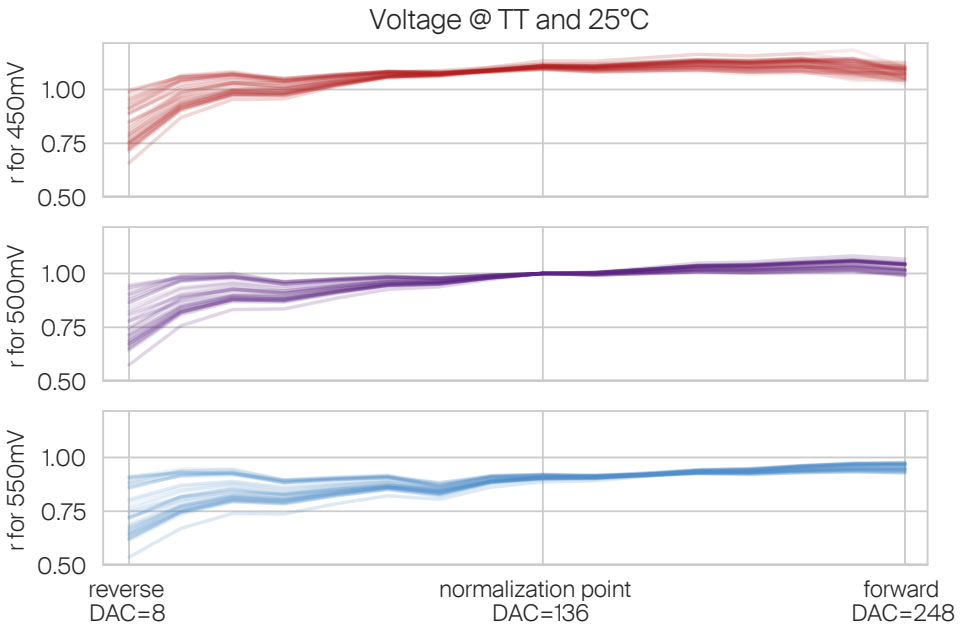



Figure 5.12 – Normalized relative frequency variation of the ring oscillators across supply voltage when sweeping the DAC code. The order is 450 mV (top/red), 500 mV (center/purple), and 550 mV (bottom/blue). An offset corresponding to $V_{DD,typ}/V_{DD}$ can be observed.

The core frequency was estimated by a sweep, running the same application performing a continuous matrix multiplication while sweeping the clock frequency up until either observing a calculation error or until JTAG connectivity is lost. Figure 5.13 plots the core frequency across corners on top of the ring data presented in the previous discussion.

The Core frequency mostly follows the ring frequencies, however it tends to follow the scaling of the less graciously scaling cells, in particular for the reverse bias condition. This behaviour is not completely unexpected: the rings consist of a single cell type, while the critical path within the core is constructed from a mix of different cells. If we consider deep sub-threshold operation we expect some transistor stacks to perform worse than others due to local variation. With the exponential slope in the sub-threshold domain the corresponding delay variations can become significant, resulting in a situation where the worst case cells dominate the delay on the critical path[1].

As a result, reverse operation requires a larger timing margin which is typically not an issue as the clock frequency can be relaxed: in most applications the core would only infrequently sample some sensor data and decide with a simplistic model whether it needs to wake up for fast processing. Hold timing margins may however have an effect on the overall system timing and power consumption.

### 5.4.4 Hardware Verification of the ADVBB Model

Finally, we use the integrated test circuitry to verify the timing model presented in Chapter 3. We use an open loop approach, where the model prediction for a constant frequency trajectory through the $V_{DD}$-$V_{BB}$-design space is applied to the actual chip by externally supplying $V_{DD}$, $V_{BN}$, and $V_{BP}$.

#### Timing Model Matching

First, we will explore the accuracy of the timing aspects of the ADVBB model presented in Chapter 3. The model presents itself as a simple look up table, providing the tuple of reference dynamic and leakage current, NMOS bias $V_{BN}$, and PMOS bias $V_{BP}$ in a coordinate system of supply voltage against reference circuit frequency. We use a separate ADVBB model for each temperature.

To do so we reuse the rings from Section 5.3 as timing monitors. We used a single die, and measured the ring frequencies across temperature for −40 °C, 25 °C, and 85 °C. The starting point for the constant frequency trajectory in the $V_{DD}$-$V_{BB}$ design space was selected to be ($V_{DD}$, $V_{BN}$, $V_{BP}$)=(0.5 V, 0.15 V, 0.18 V) at TT with 25 °C, the same point as used as a reference

[1]Note that the core frequency determination is less accurate than the ring frequency due to a) the discrete step nature of the measurement approach and b) the critical path potentially shifting which may result in different failures triggered in the different corners.

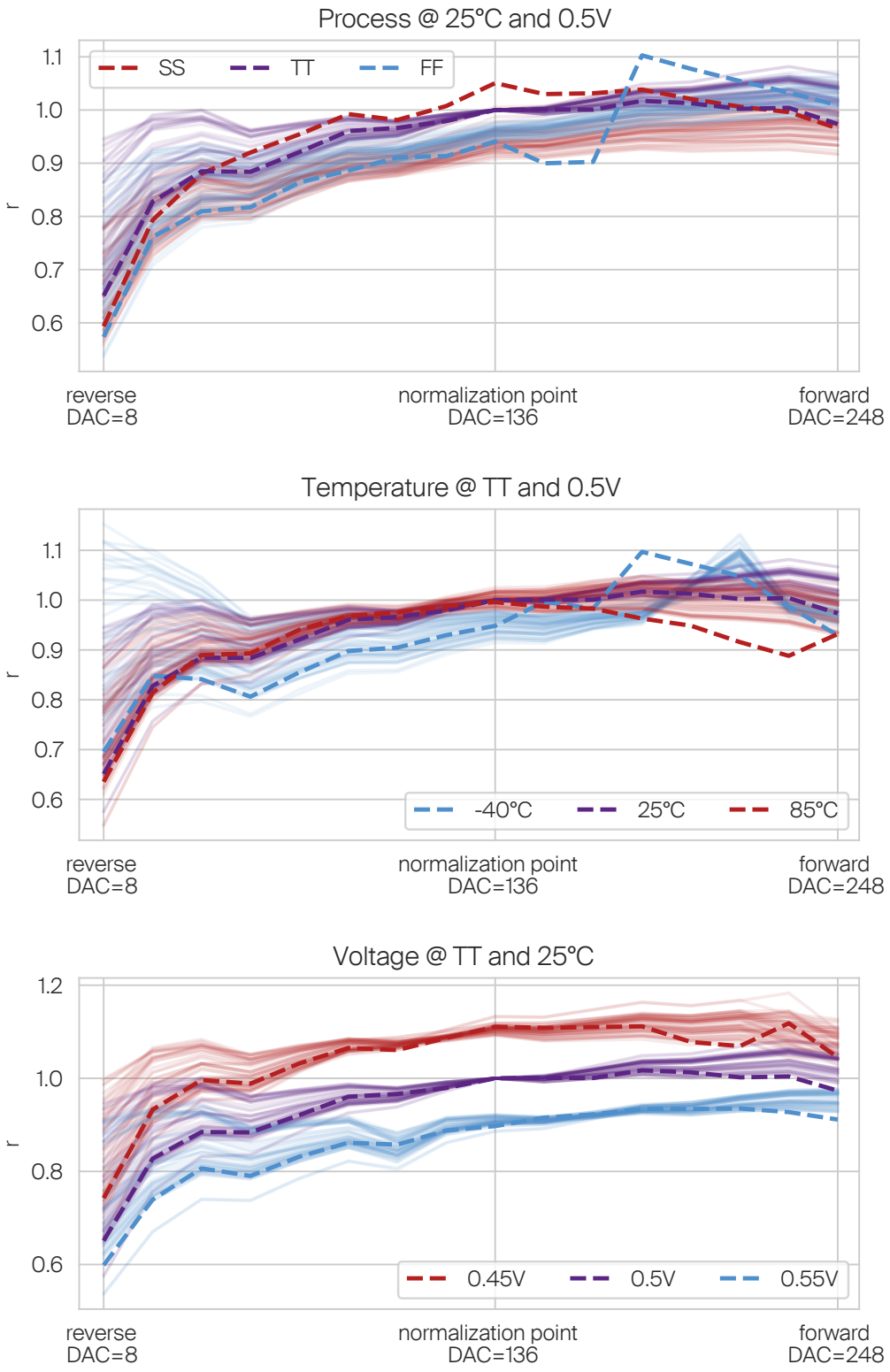



Figure 5.13 – Normalized relative frequency variation of the core (dashed) against the ring oscillators (solid) across process corners when sweeping the DAC code

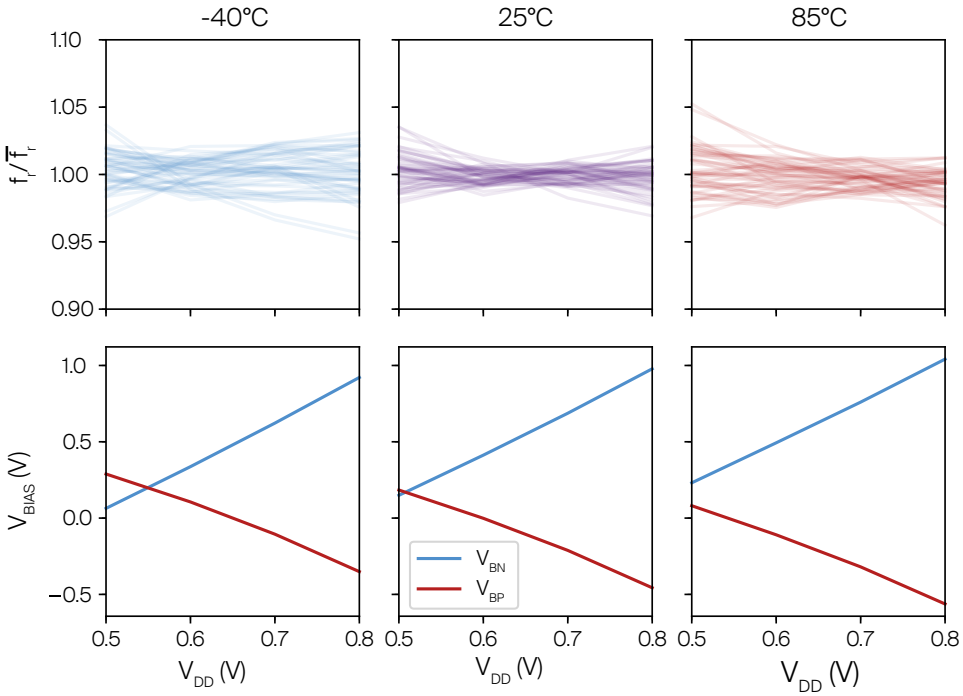



Figure 5.14 – Ring frequency $f_r$ normalised with the mean frequency $\overline{f_r}$ when sweeping a constant frequency trajectory using the supply voltage / bias trajectory from the model. Note that $\overline{f_r}$ is the mean across corner and all values along the iso-frequency trajectory.

in Chapter 2, corresponding to a slight forward bias with balanced NMOS and PMOS $I_{ON}$. We then employ the the same frequency for the −40 °C and 85 °C within their respective models to find the corresponding constant frequency $V_{DD}$-$V_{BB}$ trajectories.

In the following we normalised the ring frequencies as follows

$$f_{norm} = \frac{f_r}{\overline{f_r}}, \tag{5.10}$$

where $f_r$ corresponds to a given frequency measurement for a given ring and $\overline{f_r}$ corresponds to the average frequency measurement of the same ring across the temperature corners and across the constant frequency trajectory of the $V_{DD}$-$V_{BB}$ design space. For a perfect match, between the model and the hardware we would expect near constant frequency when sweeping the three voltages along the trajectory.

Figure 5.14 shows the result of a sweep for a typical die across temperature for −40 °C, 25 °C, and 85 °C. The graphs on top show the frequencies of each ring, normalised as described to the mean frequency across all measurements. The bottom graphs show the two bias voltages given

by the model for that specific frequency trajectory against the supply voltage.

We observe a very close match, with a frequency within a range over all the rings spanning from 95% to 105%. This is comparable to the spread between the rings with different cells we have seen in the previous section, hence the model mismatch is lower than the cell to cell variation. Further, even though the measurements were done on a typical corner lot we can assume that the actual die NMOS and PMOS performance does not to fit perfectly on the typical performance from the spice model, adding additional uncertainty. We can conclude that the model predicts the required bias for a given frequency at a given supply voltage very well as we observe near perfect compensation for the predicted bias and supply voltage settings.

#### Power Model Matching

We also verify the power predictions from the ADVBB model from Section 3.2 against actual hardware. We again use the constant frequency trajectory open loop approach to adjust the bias and supply voltage according to the model and measure leakage and dynamic power on the icyflex core of the microprocessor for the same typical die for −40 °C, 25 °C, and 85 °C.

As a first step we have to extract the scaling factors $r_{dyn}$ and $r_{leak}$ introduced in section 3.4.1, used to scale the model dynamic power to the characteristic circuit activity and the circuit area dependent model leakage respectively.

This was done using a single point calibration based on the measurement at 25 °C for 0.5 V with $V_{BN}$ of 0.15 V and $V_{BP}$ of 0.18 V. This point was selected to a) be centred in the temperature range and b) with as much leakage as possible to keep the effects of noise during measurements as little as possible.

With the leakage factor only being area dependent and the dynamic factor being constant for constant frequency we can apply the same factors across all corners.

The measured dynamic $P_{dyn}^{meas}$ and leakage power $P_{leak}^{meas}$ are now plotted in Fig. 5.15 on the top both against the ADVBB model predictions $P_{dyn}^{model}$ and $P_{leak}^{model}$ after calibration. In addition, the bottom shows the relative error $P_{dyn}^{meas}/P_{dyn}^{model}$ and $P_{leak}^{meas}/P_{leak}^{model}$.

We observe a very tight match between the model and the measured dynamic power in the order of around 5% across the corners. The leakage error observed is higher, but considering that we observe a scaling over several orders of magnitude through the constant frequency cut it is still in an acceptable range between a factor of 0.4 and 2.

### 5.4.5 System Results

This Section will covers the results of the digital system measurements, considering the achievable frequency as well as the leakage and dynamic power for the core as well as for the memory.

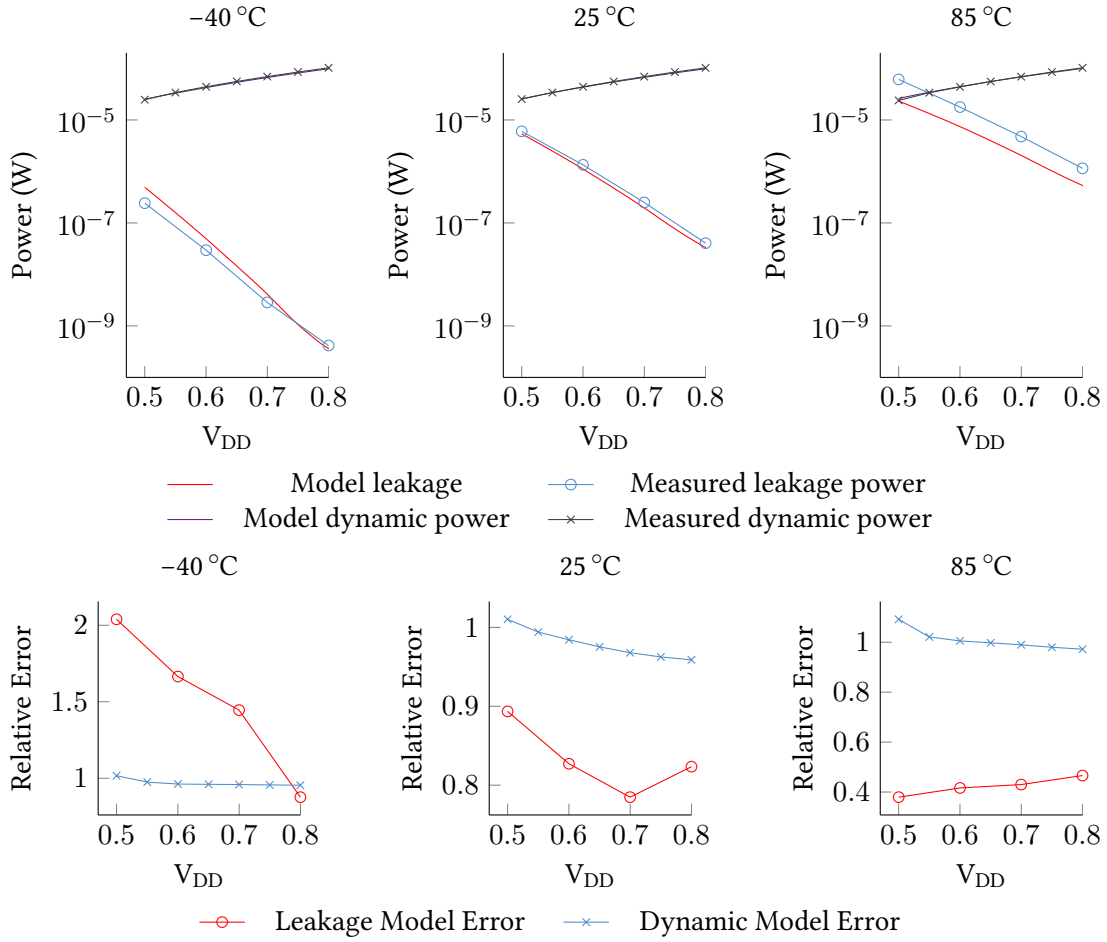



Figure 5.15 – Core power measurements in open loop biasing applying bias voltages as determined by the model for a constant frequency cut of the $V_{DD}$-$V_{BB}$ design space.

### Core Measurements

The core measurements are fairly straight forward: we are interested in the maximum frequency as well as the corresponding core dynamic power and leakage for a sweep of on-current set points (represented by the corresponding DAC codes) codes across PVT. We used the same approach as for Section 5.4.3 by increasing the frequency gradually until a test program looping a continuous matrix multiplication either showes in a calculation errors or until the JTAG connectivity is lost due to timing errors on the bus. Dynamic power was obtained by measuring the core domain power through the SMU supplying the core with a large enough integration time to average over several cycles of the matrix multiplication code. Finally, leakage was determined by disabling the clock externally.

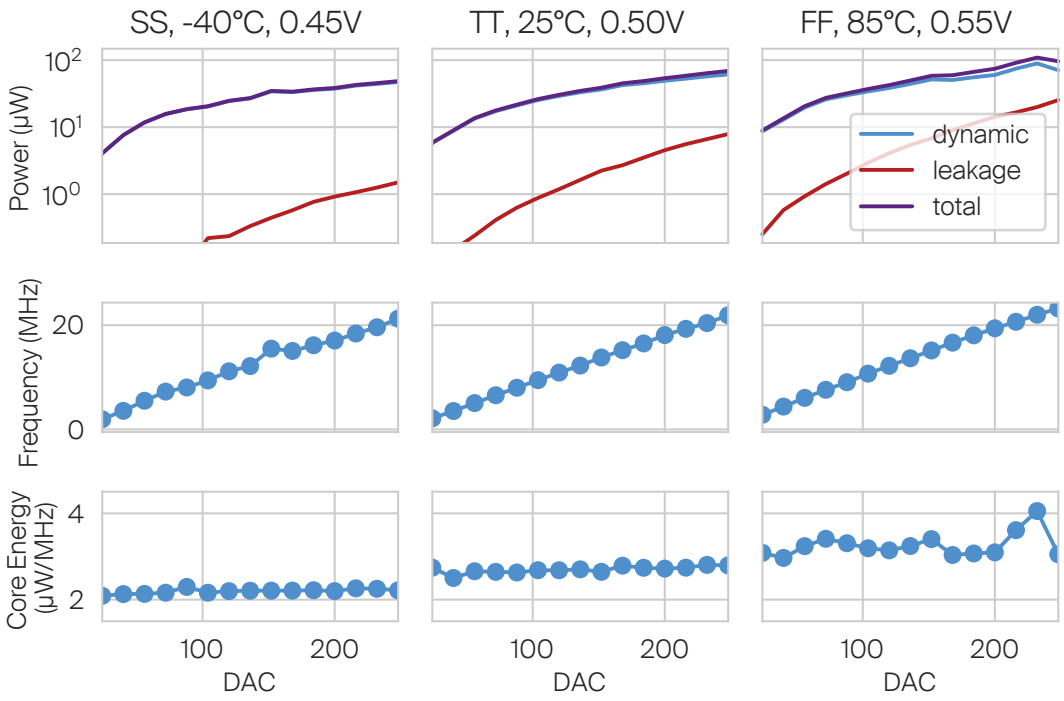



Figure 5.16 – Core power, maximum frequency, and dynamic energy across three different PVT corners.

Figure 5.16 shows the results for three PVT corners. When sweeping the on-current set point DAC code from 8 to 248 we observe, as expected, a near linear shift of the stable core operating frequency from 600 kHz to 21.2 MHz. When comparing the core frequencies across PVT we see the slight shift due to the effect of the supply voltage, but observe otherwise a very decent match.

The core power across the corners is dominated by the dynamic power, but leakage becomes no longer insignificant in the high temperature, high frequency cases due to the excessive forward bias needed to achieve the speed required.

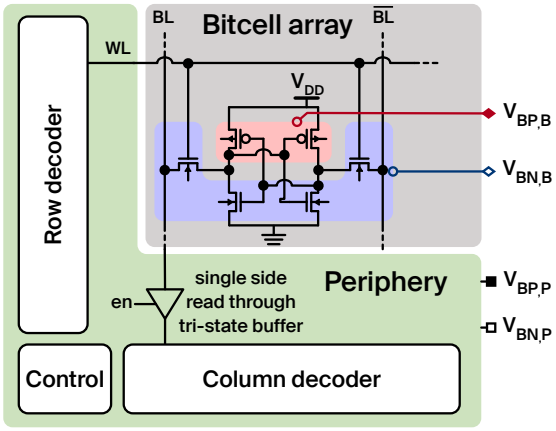



Figure 5.17 – Calanda SRAM bias architecture, using two bias domains: The array bias domain consists only out of the 6T bit cells. The periphery bias domain integrates the control, read-out buffers, row and column decoders.

Finally, the dynamic energy is nearly constant within the corner, but shifts as expected linearly with the supply voltage, ranging from 2.2 µW/MHz over 2.67 µW/MHz to 3.26 µW/MHz.

### Memory Operation and Retention

Contrary to the core, the SRAM shows very limited switching activity, resulting in a highly leakage dominated power profile. The SRAM is constructed as shown in Fig. 5.17, integrating two separate bias domains for the bit cell array and periphery, respectively, allowing for an optimised bias for each of the two domains. The foundry 6T bitcell is biased as far reverse as possible to still meet timing in order to reduce the leakage as much as possible. The periphery is smaller and hence contributes less to the overall leakage and needs to be controlled such that no timing errors occur when communicating with the core. The measurements were operated from the integrated core by writing and verifying check patterns onto the macro. Due to the vastness of the design space the measurement campaign is not capable to provide robust statistical data about the static noise margins, but sufficient to provide a general insight of the design space available to the user.

Finding the optimal bias conditions is not trivial: the four bias voltages, setting the NMOS and PMOS bias for bitcell and periphery[2] open up a four dimensional design space. We are interested in finding the operating points throughout this design space where a) the memory is functional, b) the required operating frequency is reached and c) the SRAM leakage is minimised. For a) it is necessary to write multiple test patterns into the memory over a reasonably large memory region to cover eventual variations followed by a read back and comparison in order to verify integrity. For b) all the steps from a) have to be repeated at different clock frequencies. Finally, for c) we need to measure the leakage power by clock gating the whole system and choose the

[2]The periphery is consisting out of row & column decoder as well as control.

functional point where leakage is found to be minimal.

A naïve approach to find the optimal bias combination would be a full sweep of the 4D design space, defined by the NMOS and PMOS bias voltage for bitcell array and periphery respectively. However, with the valid range of the bias spanning a range of 1.6 V even a fairly coarse sweep in 50 mV steps explodes to $32^4 = 1'048'576$ potential operating points which is clearly excessive, not even considering the need to also sweep the operating frequency in each point. However, we are only interested in operating points where the memory is functional and we can make the reasonable assumption that the region where the memory is operable is continuous.

We can exploit this assumption by considering a known working starting point for a given frequency $f_{run}$, consisting of the bias voltage pairs for the bitcell ($V_{BNB}$, $V_{BPB}$) and periphery ($V_{BNP}$, $V_{BPP}$). Such a point in our four dimensional grid can be defined by the following tuple:

$$\divideontimes_{a,b,c,d} = (V_{BNB}, V_{BPB}, V_{BNP}, V_{BPP}) \tag{5.11}$$

This point is scheduled first for measurement. We can now follow the flood fill approach sketched in Alg. 2: if, and only if, the measurement reveals a functional point, we measure all neighbour points as long as they are a) within the legal boundaries of the biasing ranges and b) have not been measured yet. The required eight direct neighbours of a functional operating point can be defined by the offset $V_{step}$ in the grid as follows:

$$\begin{aligned}
\divideontimes_{a+1,b,c,d} &= (V_{BNB} + V_{step}, V_{BPB}, V_{BNP}, V_{BPP}) && (5.12)\\
\divideontimes_{a-1,b,c,d} &= (V_{BNB} - V_{step}, V_{BPB}, V_{BNP}, V_{BPP}) && (5.13)\\
\divideontimes_{a,b+1,c,d} &= (V_{BNB}, V_{BPB} + V_{step}, V_{BNP}, V_{BPP}) && (5.14)\\
\divideontimes_{a,b-1,c,d} &= (V_{BNB}, V_{BPB} - V_{step}, V_{BNP}, V_{BPP}) && (5.15)\\
\divideontimes_{a,b,c+1,d} &= (V_{BNB}, V_{BPB}, V_{BNP} + V_{step}, V_{BPP}) && (5.16)\\
\divideontimes_{a,b,c-1,d} &= (V_{BNB}, V_{BPB}, V_{BNP} - V_{step}, V_{BPP}) && (5.17)\\
\divideontimes_{a,b,c,d+1} &= (V_{BNB}, V_{BPB}, V_{BNP}, V_{BPP} + V_{step}) && (5.18)\\
\divideontimes_{a,b,c,d-1} &= (V_{BNB}, V_{BPB}, V_{BNP}, V_{BPP} - V_{step}). && (5.19)
\end{aligned}$$

During measurement each grid point is either marked as working or failed, while also collecting leakage and dynamic power. During this process new measurement points are spawned recursively until all reachable points have been measured. The result is such that all points on the hypersurface of the four dimensional cloud are either marked as failed or reach the edge of the legal biasing range. The cloud of working points is now complete for the given frequency $f_{run}$.

**Algorithm 2** Floodfill measurement of the design space

```
function FloodFillMeasureSRAM(*a,b,c,d)
    if AlreadyMeasured(*a,b,c,d) or InvalidPoint(*a,b,c,d) then
        return
    end if
    SetBias(*a,b,c,d)
    if SuccessfulWriteAndVerifyPattern() then
        MarkValid(*a,b,c,d)
        FloodFillMeasureSRAM(*a+1,b,c,d)
        FloodFillMeasureSRAM(*a-1,b,c,d)
        FloodFillMeasureSRAM(*a,b+1,c,d)
        FloodFillMeasureSRAM(*a,b-1,c,d)
        FloodFillMeasureSRAM(*a,b,c+1,d)
        FloodFillMeasureSRAM(*a,b,c-1,d)
        FloodFillMeasureSRAM(*a,b,c,d+1)
        FloodFillMeasureSRAM(*a,b,c,d-1)
    else
        MarkFailed(*a,b,c,d)
    end if
    return
end function
```

The frequency can now be reduced and the failed points can be scheduled for re-measurement at a lower frequency. When repeating this process by continuously reducing the frequency we can map out the complete operational range with the corresponding frequencies, leakage, and dynamic power while minimising the number of measurements to the functional range.

Finally, the search can be extended to identify low power retention points: starting from a known working reverse operational point data is written to the SRAM, the chip is sent to a further reverse retention point, kept there for a few seconds, and then returned to the starting bias for reading back and verifying the data.

The results of the sweep for the 64 kB memory on the chip are depicted in Fig. 5.18. On the left we see the 2D projection of the four dimensional cloud into the $V_{BPP}$-$V_{BNP}$ and $V_{BPB}$-$V_{BNB}$ planes for the typical case. The black dots marking the lowest leakage point for that particular frequency. On the top right we see the results for the bitcell array only in the typical and two extreme cases using a FF and SS corner lot sample applying 0.55 V at 85 °C and 0.45 V at −40 °C. The black dots again represent the lowest leakage point. Below, we plot the corresponding leakage currents across frequency. The black lines correspond to the equivalent constant frequency lines plotted above across the point cloud as well.

We can now define operating points for use of the SoC: Table 5.2 presents a potential slow and

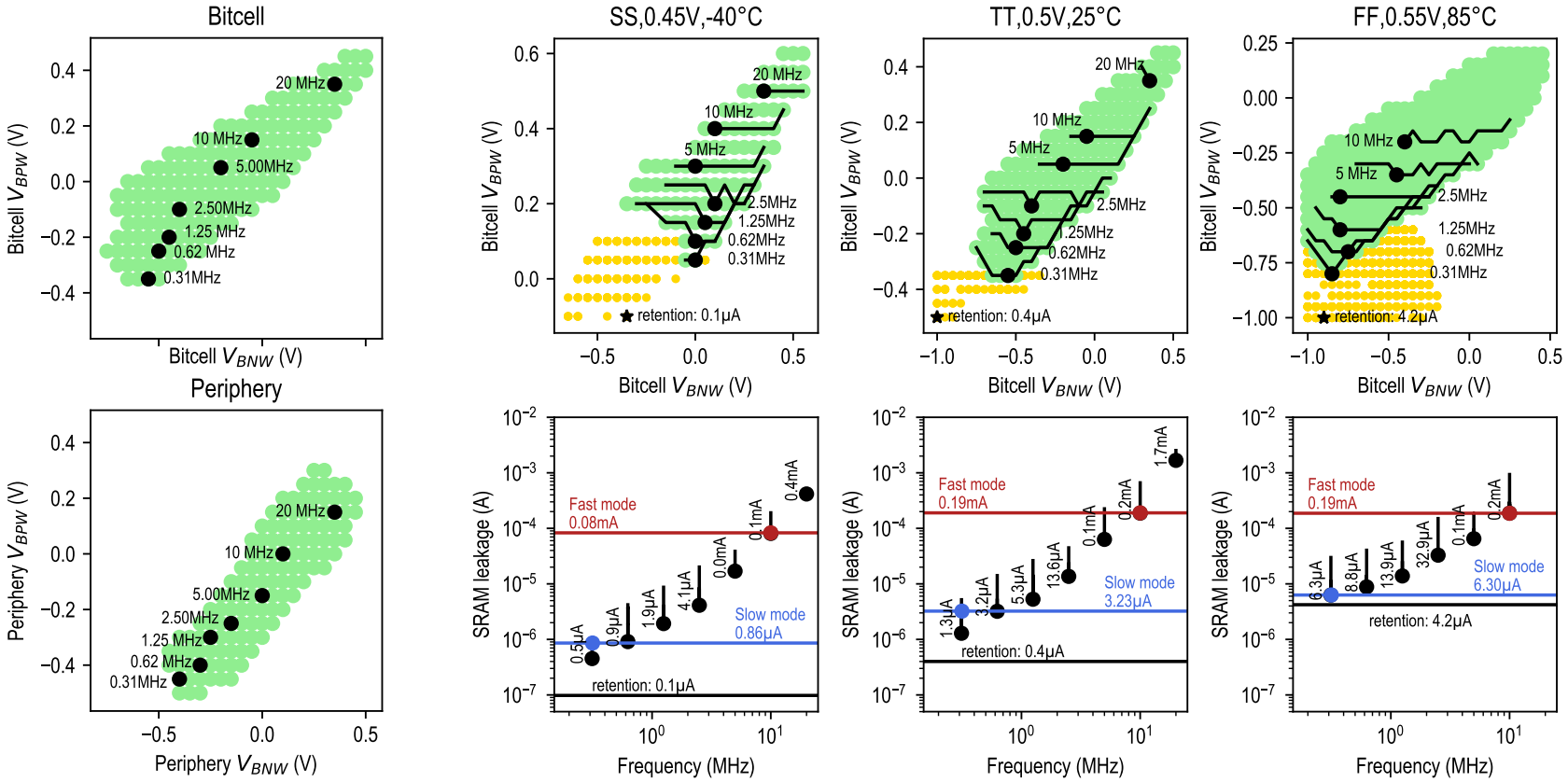



Figure 5.18 – Left: Functional SRAM bias conditions in TT 0.5 V 25 °C. The SRAM functionality is defined in a 4D bias design-space including periphery and bit-cell biases. Black dots mark minimum leakage points. Right: 64 kB SRAM measurement in extreme and typical PVT. Green dots are working ($V_{BNW}$,$V_{BPW}$) pairs. Black lines indicate constant frequency fronts. Black dots indicate lowest leakage for the given frequency line. Yellow dots are retention ($V_{BNW}$,$V_{BPW}$) configurations. The black star is the lowest retention leakage. Fast and Slow mode leakages are plotted against frequency by red and blue lines.

Table 5.2 – Measurements for Fast and Slow modes of leakage and dynamic currents for the 32 bit RISC Core as well as the SRAM in extreme and typical PVT conditions.

| | FF<br>0.55 V<br>85 °C | | TT<br>0.5 V<br>25 °C | | SS<br>0.45 V<br>−40 °C | |
|---|---|---|---|---|---|---|
| | RISC Core | | RISC Core | | RISC Core | |
| | leakage | dynamic | leakage | dynamic | leakage | dynamic |
| Slow (0.31 MHz) | 250 nA | 1.84 µA | 54 nA | 1.66 µA | 14 nA | 1.52 µA |
| Fast (10 MHz) | 4.08 µA | 59.3 µA | 1.19 µA | 53.4 µA | 0.24 µA | 48.89 µA |
| | 64 kB SRAM | | 64 kB SRAM | | 64 kB SRAM | |
| | leakage | dynamic | leakage | dynamic | leakage | dynamic |
| Retention | 4.2 µA | gated | 0.4 µA | gated | 0.1 µA | gated |
| Slow (0.31 MHz) | 6.3 µA | 1.7 µA | 3.23 µA | 1.55 µA | 0.86 µA | 1.43 µA |
| Fast (10 MHz) | 190 µA | 54.8 µA | 190 µA | 50 µA | 80 µA | 46.1 µA |

potential fast mode, selected such that the frequency is achievable across PVT for both the core and SRAM while not going too far forward where the digital system would suffer from excessive leakage. For the SRAM these points are selected such that we find a constant current DAC setting which is equal across PVT, and that the worst case corner (FF 0.55 V 85 °C) leakage is minimised. The modes are also plotted in Fig. 5.18, as the blue and red lines marked fast and slow, respectively.

We observe a leakage factor of 93.0, 58.8 and 30 between the slow and fast operation modes across the three corners for the SRAM, showing that adaptive body control is a particularly efficient method to reduce the leakage power on SRAMs. Furthermore, the SRAM leakage dominates the power consumption across all PVT cases, making the memory the prime target for system level optimisation strategies. Potential options include segmented memories with adaptive biasing of the currently not used portions of the memory, assuming that the application is using it typically in consecutive patterns which for example is the case for the instruction memories under the assumption that no excessive jumps are needed.

Note that this sweep is insufficient for proving a robust static noise margin analysis on the acceptable imbalance and should be rather seen as a general direction. A proper answer would require a much more exhaustive measurement campaign, running on more samples and repeating the same test over a larger memory space area. However, the paper presented by Misawa et al. at S3S 2017 [70] can provide some insight.

Table 5.3 places the circuit against a selection of comparable SoCs.

Table 5.3 – Comparison of the Calanda SoC with state of the art.

| | Calanda | ISSCC2018 [34] | ISSCC2018 [54] | ISSCC2014 [71] | ISSCC2015 [13] | JSSCC2017 [55] |
|---|---|---|---|---|---|---|
| Technology | 55 nm DDC | 28 nm FDSOI | 180 nm Bulk | 65 nm Bulk | 65 nm Bulk | 40 nm Bulk |
| Core | 32 bit RISC | LVT core | 16 bit MSP430 | n/a | 32 bit Cortex M0+ | 32 bit Cortex M0 |
| Core dyn (µW/MHz) | 2.67 | n/a | 14 | n/a | 11.7 | 8.8 |
| Core leak (nW) | 27 | n/a | <1 | n/a | 20 | n/a |
| SRAM (kB) | 64 ULL 6T | n/a | 2 SCM | 16 XLL 6T | 8 LV 10T | 64 6T |
| SRAM dyn (µW/MHz) | 2.5 | n/a | n/a | 25 | n/a | n/a |
| SRAM retention (nW/kB) | 3.13 | n/a | n/a | 0.26 | 15 | n/a |
| SRAM bitcell size ($\mu m^2$) | 0.425 | n/a | n/a | 2.159 | 3.64 | n/a |
| Retention (Core + 4kB SRAM) (nW) | 39.52 | n/a | n/a | n/a | 80 | n/a |
| Frequency (MHz) | 0.31-10 | 9 MHz target | NM:0.016-2.8<br>LSM: 1 Hz-4 Hz | 142.9 | 0.029-66 | 0.8-50 |
| Frequency variation in PVT (%) | Fast: ±6%<br>Slow: ±21% | ±3.5% | Only Die-2-Die process 3$\sigma$<br>NM: ±23%<br>LSM: ±36% | n/a | n/a | n/a |
| Process measured. | SS, TT, FF | Single process | Single process | Single process | Single process | Single process |
| Supply range | 0.5 V±10% | 0.35-1 V | 0.2 V-1.1 V | 1.2 V | 0.19 V-1.2 V | 0.2 V-0.5 V Core<br>0.6V SRAM |
| Bias range. | $-1\,V<V_{BB}<0.6\,V$ | NW: 0 to 1.8 V<br>PW: -1.5 to 0 V | n/a | $-0.3\,V<V_{BB}<-0.1\,V$ | n/a | n/a |
| Temperature range | -40 to 85 °C | -40 to 125 °C | 0 to 45 °C | 25 to 125 °C | 25 to 70 °C | 0 to 70 °C |

## 5.5 Conclusion

Calanda has shown that the general idea of the biasing concept based on matched on-currents is a feasible method to achieve a matched biasing for variability compensation.

We have shown that the corners both in temperature as well as in process are well compensated, with a representative set of standard cell based ring oscillators that turn out to be closely matched to the core frequency. Furthermore we have shown that the on-current approach alone is overcompensating the circuit speed for supply voltage variation.

We were able to verify the model extracted from the standard cell characterisation experiments, achieving near perfect frequency matching when following the predicted constant frequency trajectory. The power prediction from the model is near perfect for the dynamic power component while following the general shape of the leakage curve, albeit with a bit of acceptable offset, considering the steep slope.

An exhaustive measurement exploration of the design space has shown that adaptive biasing of the SRAM has the potential to reduce the leakage during retention by a factor of more than 45 in the worst case and 475 in the typical case, promising significant leakage gains. The results suggests that memory banking should be implemented in order to keep the currently unused memory in a low leakage retention state whenever possible.

# 6 Nakayama: Analog On-Current Regulation with Secondary FLL based Regulation Loop

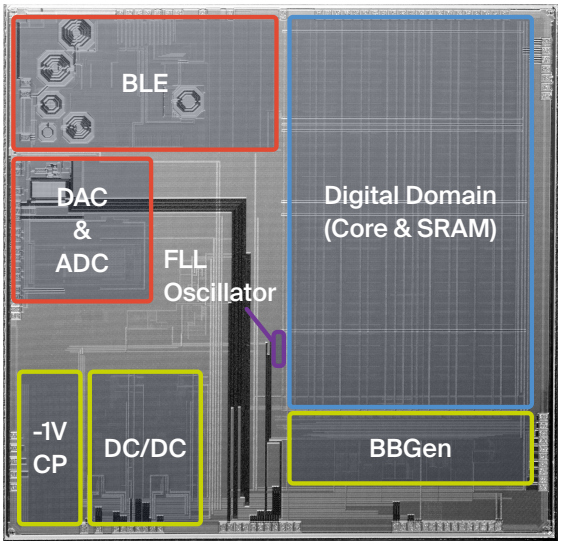



Figure 6.1 – Annotated die photo of the Nakayama SoC.

With Calanda we have seen that the on-current based biasing approach works well, with the minor drawback of a residual voltage induced variation of the propagation delay which is not compensated. Further, by relying on an external clock, Calanda can not directly adjust its bias to the clock, adding an additional burden on the user.

Nakayama addresses these drawbacks by extending the biasing concept presented for the Calanda chip with a secondary control loop, that adjusts the on-current with reference to the frequency of an on chip ring oscillator to keep it locked to an external reference clock. This approach provides the developer with a meaningful frequency knob while also overcoming the supply voltage variation aspect.

We have also seen that body control on the SRAM can be a very effective tool to reduce leakage. However, Calanda had only a single monolithic SRAM bias block which limits the options for the user when designing power saving modes. Nakayama tackles this issue with a segmented

SRAM, offering a fine grained and independent bias control for different parts of its memory.

## 6.1 System Architecture

Figure 6.2 shows a high level block diagram of the Nakayama SoC, developed within CSEM. The SoC is intended for IoT applications, integrating an IcyFlex-V 32 bit RISC-V core [72], [73] together with a Bluetooth LE transceiver [74], an ADC and a DAC, as well as typical micro-controller peripherals such as SPI, UART, and GPIOs. On chip voltage regulators produce the supplies for the IP cores, bias system, as well as for the core.

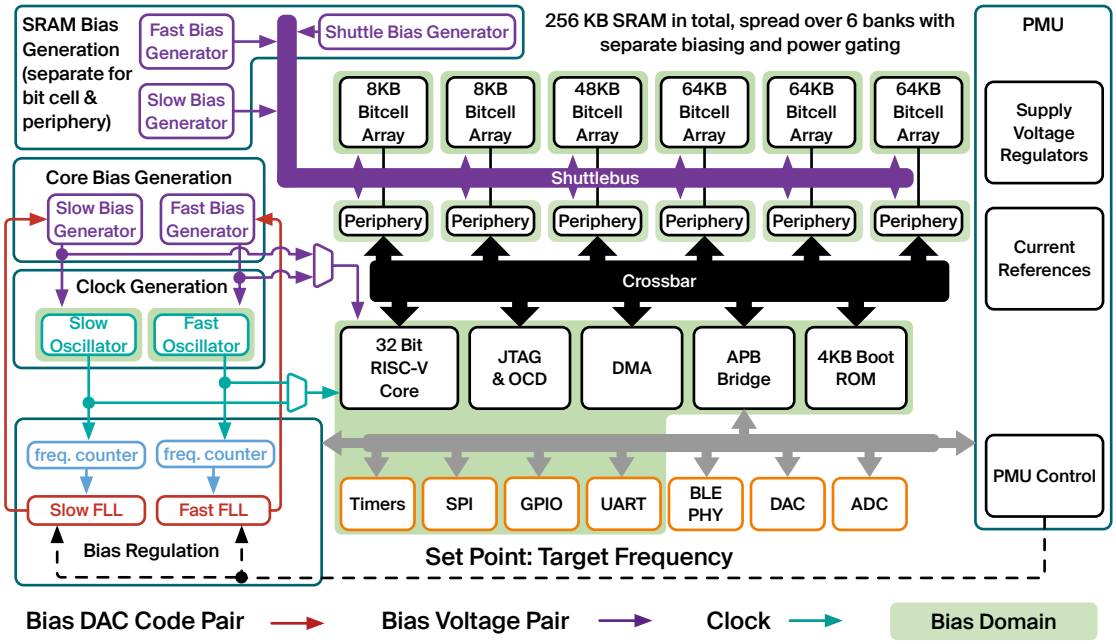



Figure 6.2 – Nakayama system overview: a 32 bit RISC-V core with 256KB SRAM, standard microcontroller periphery, ADC, DAC, BLE radio, PMU with on chip regulators as well as an adaptive body biasing system.

### 6.1.1 Core Bias Subsystem

The core can be fed by two bias generators, a forward bias oriented fast bias generator and a reverse oriented slow bias generator for implementing an operation and retention mode, respectively. Accordingly two identical oscillator instances are continuously running, with their oscillation periods set by the fast and slow bias generators respectively. The frequencies generated by the oscillators are used as system clock sources and as timing reference for the feedback loop. To this end, the frequencies are determined with two frequency counters each of which are then fed into an FLL which adjusts the DAC codes, i.e. the on-current set points, for the bias generators such that the oscillator frequency is equal to the frequency target set by the user. This parallel approach with two bias generators has the benefit that the regulation both

for retention and operation is continuously running, allowing for a fast mode switch without the need of waiting for the slow FLL regulation. For the core, the clock and bias source can be selected with software, allowing the user to define the operating modes for their particular application.

### 6.1.2 Fine Grained Memory Bias Subsystem

As we have seen in Tab. 5.2 for Calanda, the memory contributes a majority of the power to the overall system - in particular if the chip operates in a forward bias condition for high frequency operation through the increase in leakage. However, with the system in retention mode we can expect only small memory requirements for most applications (e.g. sampling some sensor data, waiting for and processing interrupts). Furthermore, in a forward operation scenario, we might need some large scratchpad memory for evaluating the user algorithm, holding intermediate results which will not be reused after evaluation finished. This scratchpad can be power gated safely when returning back to retention mode while other banks, containing algorithm parameters, should be kept available for the next evaluation, requiring the use of reverse bias to keep the corresponding leakage penalty as low as possible.

Hence, from a user perspective, a flexible system is desirable where only the sections of memory that are currently in use are in a forward biased condition, while the remaining blocks remain in a power-off or retention state. Consequently, the Nakayama SoC partitions its 256 kB of SRAM into six banks of staggered sizes. The first two banks cover only 8 kB each, with the intention of being available both during retention and operation. The remaining memory consists of one 48 kB bank and three 64 kB banks, large enough to store and evaluate a small neural network, implementing keyword spotting [75], simple image recognition tasks [3], [76], or movement classification for wearables [77]. Each bank has a power gate and its biasing can be selected through an analog cross bar ("Shuttlebus") between the output of a retention, shuttle, and operation body bias generator. This approach allows to define the memory power management in software, providing the desired fine grained control to the user.

## 6.2 Biasing Concept Details

The Nakayama bias system, depicted on the right of Fig. 6.3 against the previous Calanda system on the left. Nakayama employs an extended version of the previous design by adding a secondary frequency control loop. Calanda allowed to set the bias voltages only by adjusting the on-current, but, as we have shown in Section 5.4.3, there is a remaining effect on circuit delay due to the supply voltage variation which is not considered by just setting the on-current. Hence, Nakayama extends the on-current based biasing system with a secondary FLL based control loop, that measures the offset in delay and adjusts the DAC codes for on-current control accordingly. This secondary control loop further allows to automatically adjust the bias for a given frequency set point instead of a current set point that must manually be related to the corresponding operating frequency.

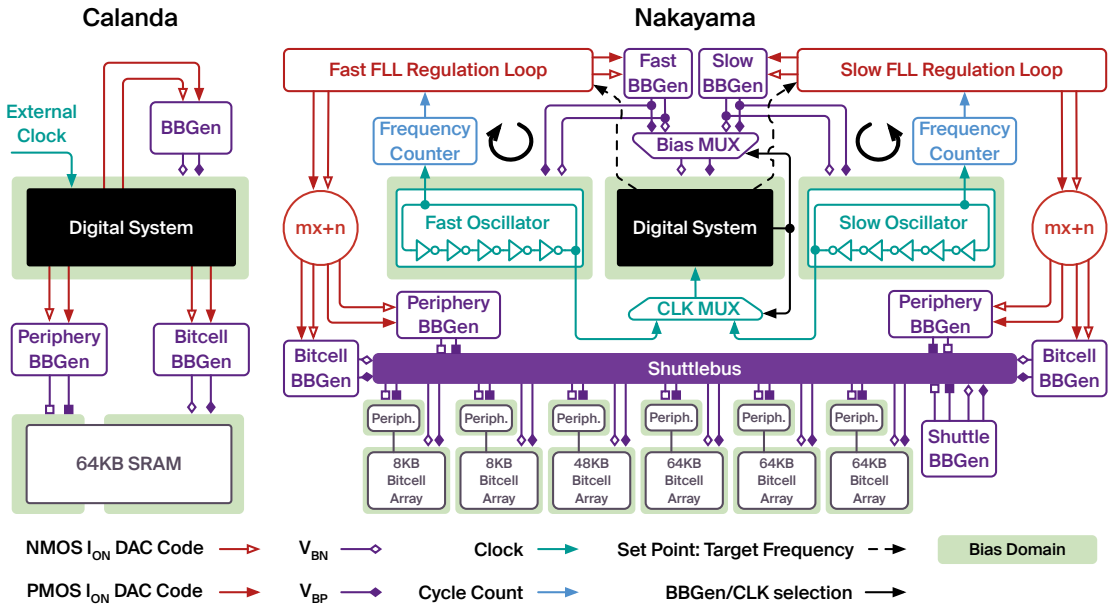



Figure 6.3 – Comparison of biasing systems: Calanda on the left, Nakayama on the right. Calanda supplies the bias of the core directly with the DAC values for the bias generator. Nakayama adds two secondary control loops, where the core sets a target frequency for fast operation and retention/slow operation. The two control loop try to achieve the target frequency by adjusting the DAC values for the BBGen, resulting in a shift of oscillator period. Further, The control outputs also set the operation and retention points for the SRAM with the possibility of a linear transformation.

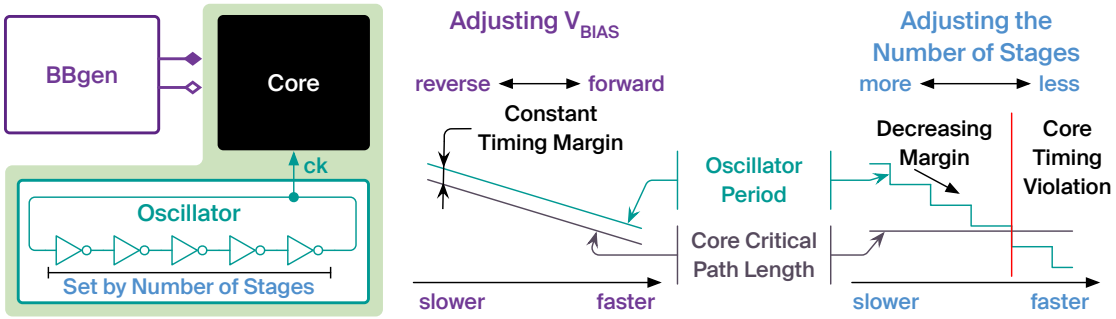



Figure 6.4 – Two principles can be used to adjust the the oscillator speed: controling the bias influences the whole system speed, setting the number of stages adjusts the oscillator speed relative to the digital circuits critical path length.

As shown in Fig. 6.3 we now implement two independent biasing systems, with the bias generators (refered to as BBgen in the figure) operating identically to the Calanda design. Hence, as already shown in Fig. 5.3, we are setting the on current through a replica transistor relative to a reference provided by a current DAC. Instead of the 8 bit current-DACs used as a reference for Calanda covering the whole bias range, two separate 5 bit current-DACs were integrated. One of the two, referred to as *fast-DAC*, covers the nominal operation in a more forward range while the second DAC, referred to as *slow-DAC*, is intended for retention and slow operation.

The BBGen is provided with the DAC-values by the secondary control loop. This loop consists of a standard cell based programmable length oscillator biased identically to the core, and a frequency counter counting cycles relative to a 32.768 kHz reference clock which is compared by an FLL integrating controller to a set value. The DAC codes for the bias generators driving the oscillator wells are adjusted such that the user programmed target frequency defined by the set value is reached.

Counterintuitively, the oscillator length is thereby not used for frequency control, but rather as a knob to allow the user to add or remove timing margin for the core. This is illustrated in Fig. 6.4: the effect of an adjustment of the bias applies to both the oscillator and the core, shifting both the critical path delay of the core and the oscillator period identically. Changing the length of of the oscillator keeps the core critical path length constant, resulting in an adjustment of the oscillator speed *relative* to the critical path length of the core. Consequently, when combining both regulation knobs, an increase in oscillator delay while keeping the target frequency constant will result in a regulation towards a more forward biased circuit with added margin while decreasing the length will reverse bias the core reducing the margin its critical path.

Further, the FLL DAC outputs are also used to adjust the SRAM bias for fast operation and retention mode. The DAC codes are fed into a linear transformation unit (refered to as mx+n in Fig. 6.3), allowing to add an offset and to adjust the slope separately for the bitcell and periphery.

The SRAM blocks are connected through an analog crossbar (referred to as shuttlebus) towards the fast and slow bias generators, allowing to choose a retention or operation bias for each of the six SRAM segments. An additional shuttle bias generator is attached to the shuttle bus, with the purpose of moving a block from forward to retention and back gracefully without disturbing other blocks remaining in forward or reverse condition.

## 6.3 Oscillator Design

The biased oscillator used in the FLL of Nakayama was designed based on standard cells with the intention to have it closely follow the bias dpendent delay scaling behaviour of the digital domain. The concept, depicted in Fig. 6.5 is based on 6 stages, implementing a delay of 1, 2, 4, 8, 16, and 32 unit delays $\delta$ respectively which can be switched in and out based on selection inputs. A NAND-gate allows to stop the clock and to implement the inversion required for the oscillation.

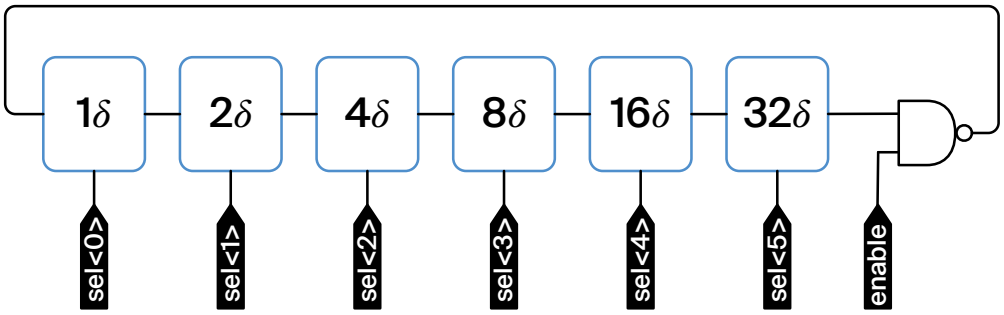



Figure 6.5 – Binary programmable length oscillator

The detailed architecture of an oscillator stage is depicted in Fig. 6.6. The unit delay $\delta$ is based on AND2-gates which proved to scale fairly similarly to the processor core in the standard cell oscillator experiments of Calanda (compare Section 5.4.3). At the same time the AND-based delay stage also allows for trivial silencing of the whole branch with the selection signal of the MUX, saving system power on the inactive branches.

A pair of inverting MUXes is used to drive the output net. The MUXes see opposite selection signals and the inputs of both are reversed. This configuration drives the output nets once through each MUX branch and hence mostly equalises the two branches, reducing the difference to only the delay of the AND-gates. Figure 6.7 illustrates the need for the double MUXes: with just a single one (red) we observe a lot of variation of the step sizes due to asymmetries in the schematic of the MUXes which causes selection depending delays with extremes ranging from 1.7 ns to 4.8 ns.

By replicating the MUX we reduce the variation significantly, with all steps falling into a range of 1.3 ns to 2.9 ns (blue). The step accuracy can be further improved by adding an additional dummy AND-gate into the loop after each MUX that is always traversed resulting in a near

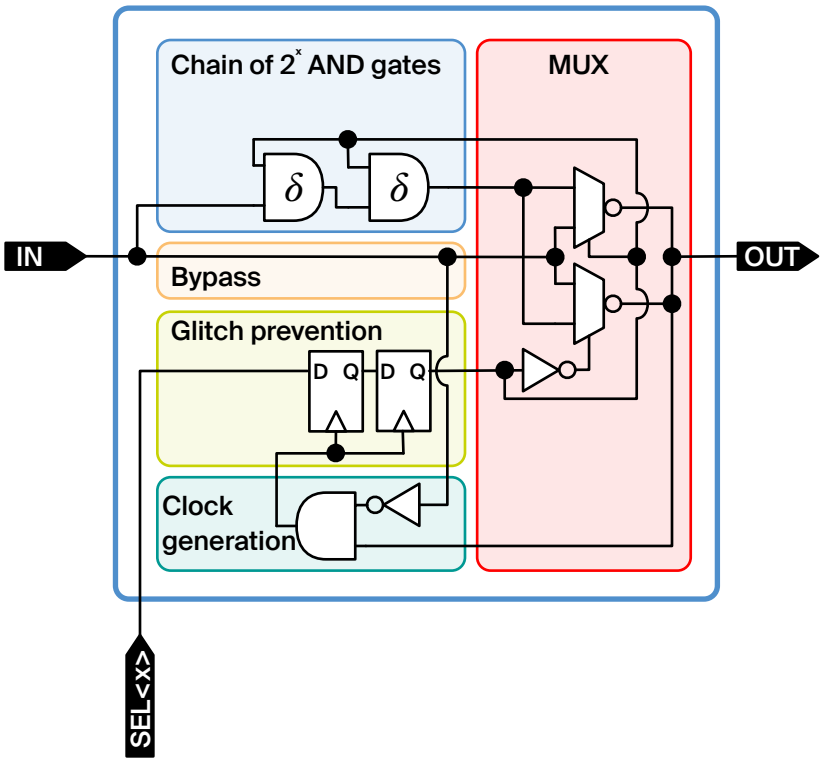



Figure 6.6 – Oscillator segment, muxing either the delay chain or the input to the output.

perfect range from 2.2 ns to 2.5 ns, however this improvement comes with a severe reduction of the maximum frequency reachable by the loop and hence has been omitted.

An important property of a programmable oscillator is its ability to be reprogrammed without injecting a) additional clock edges into the oscillation loop or b) glitches which could propagate into logic clocked by the oscillator, resulting in timing errors. In order to guarantee this property the selection is supplied through a flop, clocked by a locally generated clock, reprogramming each selection bit at a safe point in time where a flip of the selection bit will not produce a change of the MUX output. The problem breaks down into two cases: changing from the bypath path selection to the delay path and vice versa.

#### Case 1: Path Selection rising

The local clock is only generated when the input `IN` is low. The previous state of a low path selection `SEL<x>` ensures that all delay AND gates have been pulled to a low output. Besides silencing unused delay chains (to reduce dynamic power) this ensures that there is no need to wait for an edge to traverse through the delay cells, i.e. no glitch can occur.

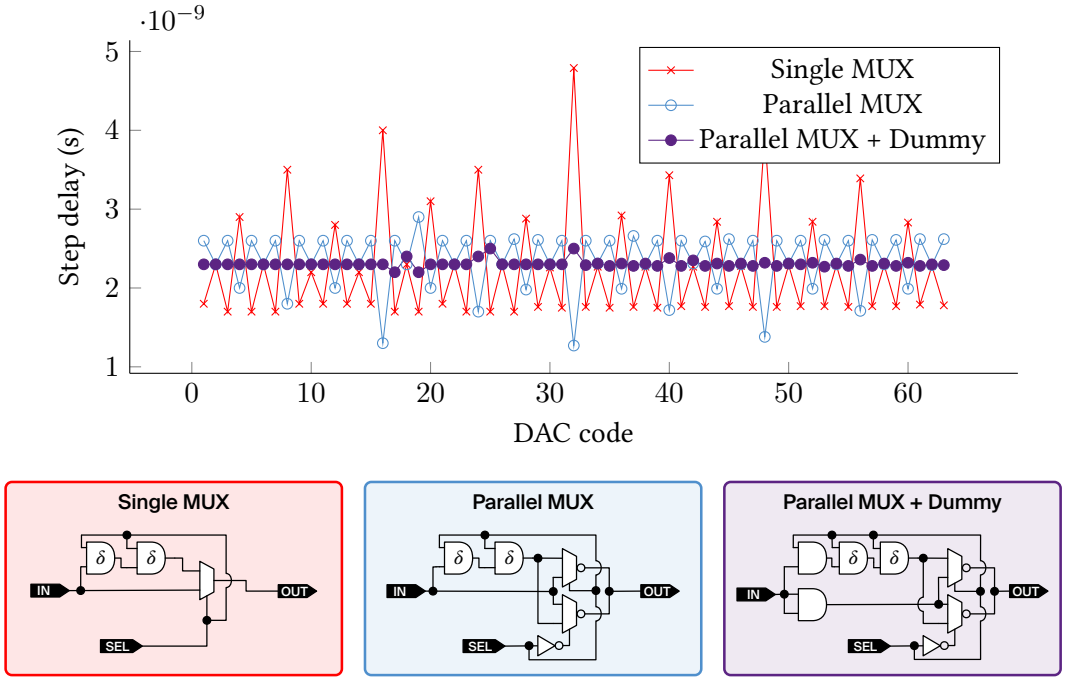



Figure 6.7 – Oscillator MUX delay step variation for three different architectures.

### Case 2: Path Selection falling

The clock again is only generated when the input IN is low. In this case, the clock signal has been generated by traversal of the whole delay chain, hence we can be certain that both inputs to the MUX are low and no glitch can occur.

A positive side effect of the glitch free reprogramming through the locally generated clock is the prevention of the MUXes driving the output net in opposite directions. This is due to the clock being generated just after the edge traversing the MUX. At that point in time we are guaranteed that both data inputs of the MUX are at the same value. The slight delay due to the inverter in front of one of the MUXes becomes negligible: the parallel MUXes will never drive in opposite directions. The flop sampling SEL<x> is doubled up in order to resolve potential metastability issues due to the alignment difference between the generated local clocks and the SEL<x> signal driven relative to the system clock.

Finally, with the inverting MUX, consecutive stages do generate the local clock at opposing phases which could result in a single shorter period when transitioning from one period to an other. This is prevented on a higher level in the digital domain through an FSM state first programming the slowest possible clock before programming a new value.

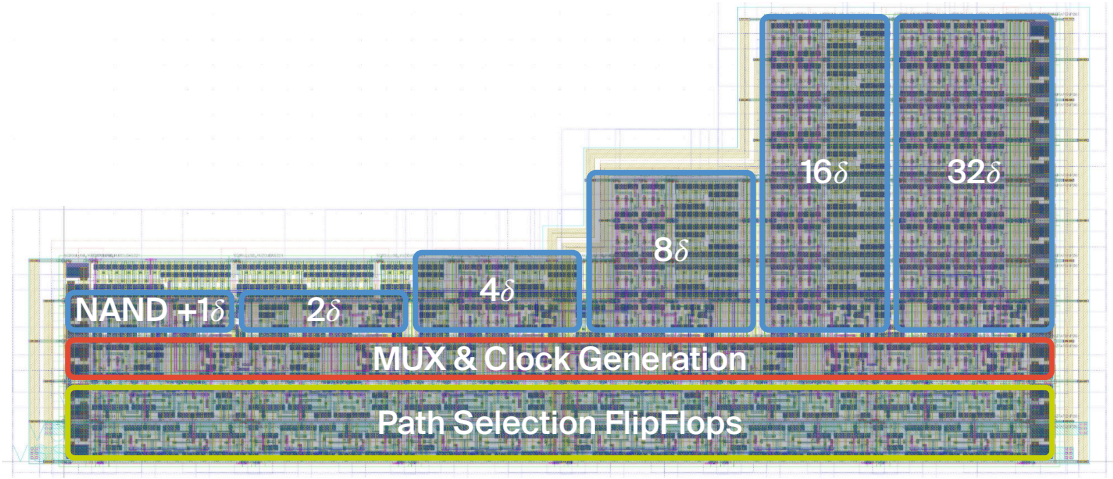



Figure 6.8 – Layout of the oscillator

## 6.4 FLL Design

The FLL implements the frequency feedback to the biasing system. A counter counts the edges of the oscillator clock during a configureable amount of reference clock cycles [78], [79]. This approach has been chosen to keep the control simple - just a simple counter is needed and comparison against a preprogrammed number is cheap.

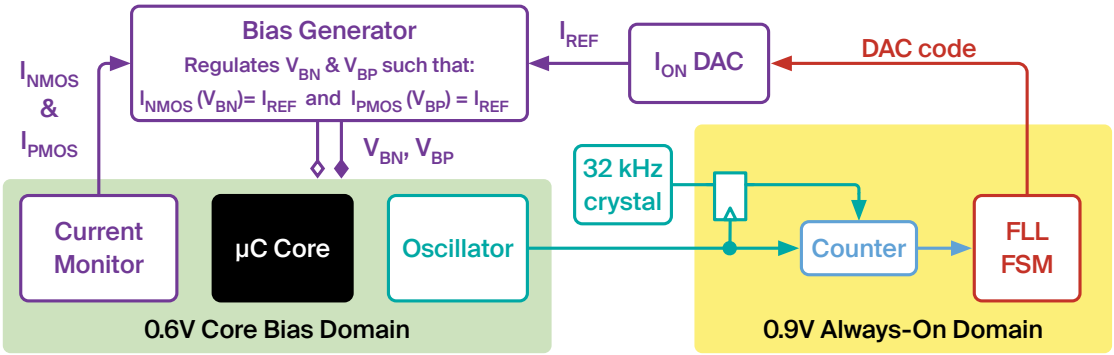



Figure 6.9 – Full body regulation loop: The system clock produced by the oscillator is counted over several cycles of the 32 kHz reference clock. The FLL regulation loop adjusts the DAC code in order to speed up/slow down the core bias domain as needed to fulfil the regulation goal. This results in a shift of the reference current which then is resulting in an adjustment of $V_{BN}$ and $V_{BP}$ as needed. This in turn changes the oscillator frequency in the direction requested by the regulator.

The regulation loop is running with the speed of the oscillator to keep regulation times low. It is placed in an always on domain without biasing and its timing is closed using a MMMC approach in order to guarantee operatability across PVT. The whole design operates under the assumption that the oscillator clock is significantly faster than the reference clock, allowing to

sample the reference clock with the oscillator clock.

### Regulation

The structure of the integrating controller of the FLL is shown in Fig. 6.10. An error signal is created by subtracting the number of expected cycles from the counted cycles and is then normalised with the number of reference cycles. As the reference cycles are expressed as powers of two, this normalisation step can be implemented with a shift. The error is then capped to a maximum of one DAC code step in order to prevent high frequency oscillations. As the DAC is only five bits wide this approach does not result in an excessive regulation speed penalty. The saturated error is integrated over time and again saturated to keep the MSB values both positive and within the DAC range. The truncated output of the integration register is directly used as the DAC code.

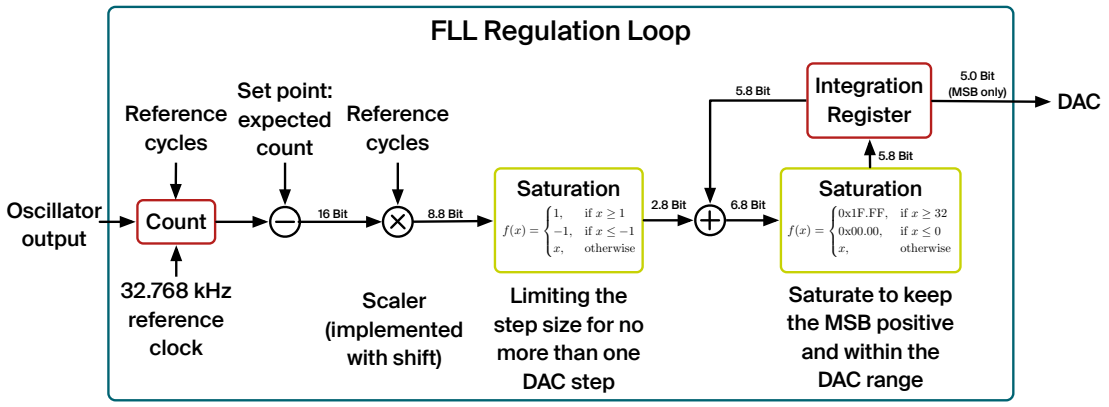



Figure 6.10 – FLL regulation loop

Figure 6.11 shows the regulation of the FLL from a mixed signal simulation together with the full analog body bias generator. The graph on the top shows the target and actual oscillator frequency over time, the bottom graph shows both the current DAC code as well as the value of the integration register with the fractional part.

The FLL is first programmed to reach a target frequency of 10 MHz which is achieved after roughly 35 cycles of the 32 kHz reference clock. We then adjust the frequency target to 15 MHz which is reached after an other 30 cycles. Finally we change the target down to 12 MHz which is achieved within 15 cycles of the reference clock. We can observe that the integration register still contains a fractional residue which is slowly decaying for a few cycles until it reaches nearly the DAC code value at which point we start seeing short oscillations between two DAC codes. Convergence is typically relatively slow—in the order of milliseconds—due to the use of the 32 kHz clock as the time reference. However, this is not an issue as rapid changes between retention and operation modes are anyway perfrormed by switching between the fast and slow bias generator and their corresponding oscillators rather than programming a different frequency. If rapid transitions between different settings of the same bias generator are required,

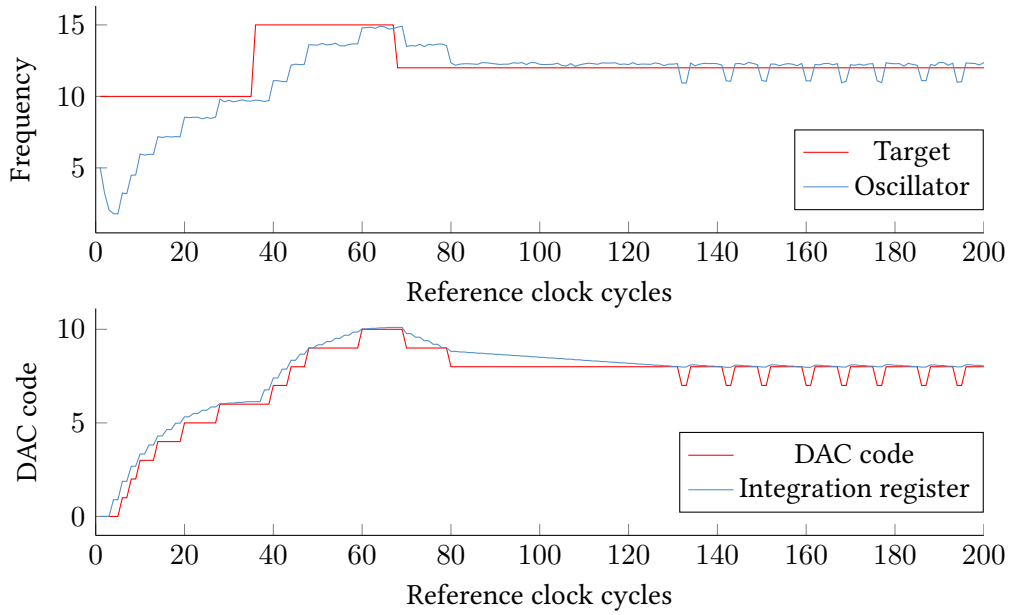



Figure 6.11 – FLL Simulation. Top: Target and oscillator frequency in MHz. Bottom: Integration register value (blue) and the DAC value derived from the MSB of the integration register.

these can be achieved by bypassing the regulation. This bypass allows to employ a look up table during mode changes to speed up the convergence before reenabling the regulation.

The oscillations around the target frequency caused by the coarse 32 bit DAC would generally not be acceptable for an FLL in a digital system as the clock frequency generated and the digital system would be typically decoupled. Hence, controlling only the average frequency over time would result in timing violations if insufficient margin is applied or the circuit would operate in sub-optimal conditions for a share of the time. This can be overcome if the voltage is scaled adaptively, similar to the voltage dithering approach presented in [80].

In our case however the core critical path delay is inherently coupled through the same bias to the oscillator speed. Hence, the core delay is follows the oscillator which makes this kind of operation mode acceptable. If a dithered frequency is not acceptable for the user, for example due to hard real time requirements or when communicating with external circuitry, the regulation can be run only occasionally in order to allow the circuit to adapt to slow environmental changes such as temperature or battery voltage fluctuations. This can be easily implemented in software, by disabling the regulation and setting the last regulation setting as a fixed value.

## 6.5 Chip Measurements

In this Section we present measurement data of the Nakayama SoC. We first characterise the oscillator as the fundamental block of the frequency control loop. We then show the operation of the FLL with a target frequency sweep, showing the generated bias voltages as well as an

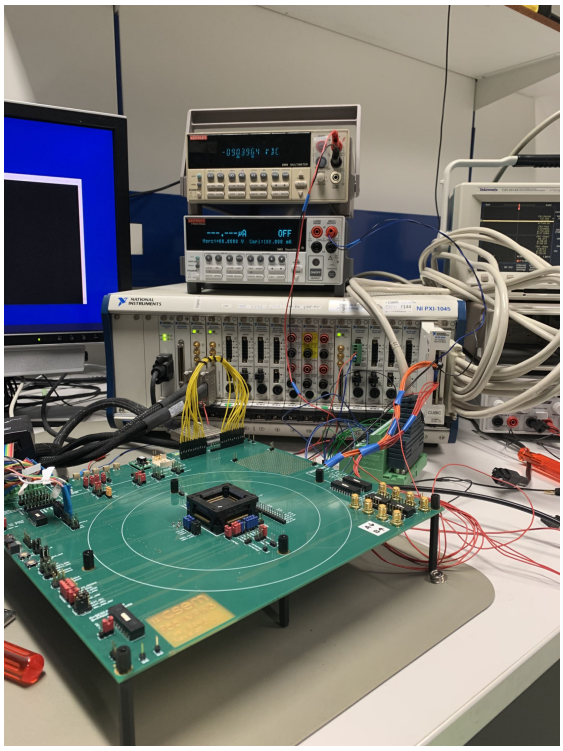
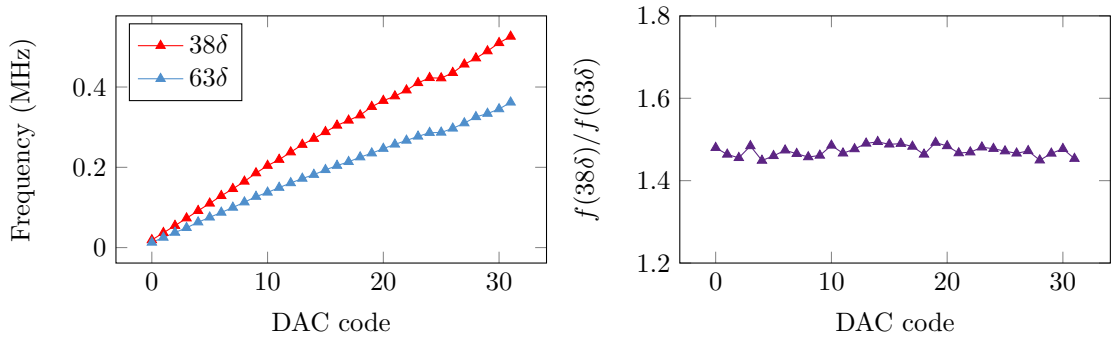


Figure 6.12 – Nakayama measurement setup.

analysis of the generated DAC codes. We finish with a system power analysis of the biasing system in operation, showing a switch from retention to operation and back.

### 6.5.1 Measurement Setup

The system measurement setup, depicted in Fig. 6.12, follows the approach used for Calanda, however has been simplified as the bias voltages and core supplies are generated on-chip. The RISC-V is now interfaced through a Segger J-Link with an updated version of OpenOCD supporting RV32 debugging, exposing the same direct memory access to the test scripts as for Calanda. The clock frequency generated on chip has been captured using a Tectronix TDS3014B oscilloscope. Static system currents have been captured using a Keithley 2400 SMU while a separate setup was used to obtain dynamic traces using a Keysight CX3300 series device current waveform analyzer. The matrix switchbox was used again with a Keithly 2000 multimeter for external supervision of the bias voltages generated by the Nakayama biasing system for the different bias domains of the SoC. The flexibility of this setup proved very useful as a large share of the system measurements have been done remotely during the Covid lockdown where access to the lab was limited.




| Length | | DAC Code | Fast Oscillator | Slow Oscillator |
|---|---|---|---|---|
| $38\delta$ | fmax | 31 | 17.98 MHz | 525.9 kHz |
| | fmin | 0 | 575.1 kHz | 19.0 kHz |
| $63\delta$ | fmax | 31 | 11.93 MHz | 361.8 kHz |
| | fmin | 0 | 287.7 kHz | 12.8 kHz |

Figure 6.13 – Top left: Slow oscillator frequency when sweeping the DAC code for the length corresponding to the core critical path with some margin as well as for the maximum length. Top right: Ratio between the two oscillator length configurations. Bottom: Range for both the slow and fast oscillator.

### 6.5.2 Oscillator Characterisation

Figure 6.13 shows the frequency behaviour of the slow oscillator when sweeping the DAC code of the corresponding bias generator across the full range for the oscillator for an oscillator length of 63 and 38, corresponding to the maximum length and the length matched to the core respectively.

When observing the slow oscillator we can observe that the programmed length, intended for margining behaves as expected mostly as expected for $^{63}/_{38}$ length we achieve slightly less than the factor of 1.66 in frequency which we would expect by just taking the ratio of delay cells due to the constant delay added to the loop by the MUXes and the additional NAND gate.

Both oscillators are identical in construction and layout, with the only difference of the bias applied for a programmed DAC value. The resulting difference in range is reported as well in Fig. 6.13.

### 6.5.3 FLL Characterisation

In the following we will analyse the FLL behaviour: Figure 6.14 shows a sweep of target frequencies on the fast FLL with the average frequency, plotting the ratio of the set frequency and the measured average frequency:

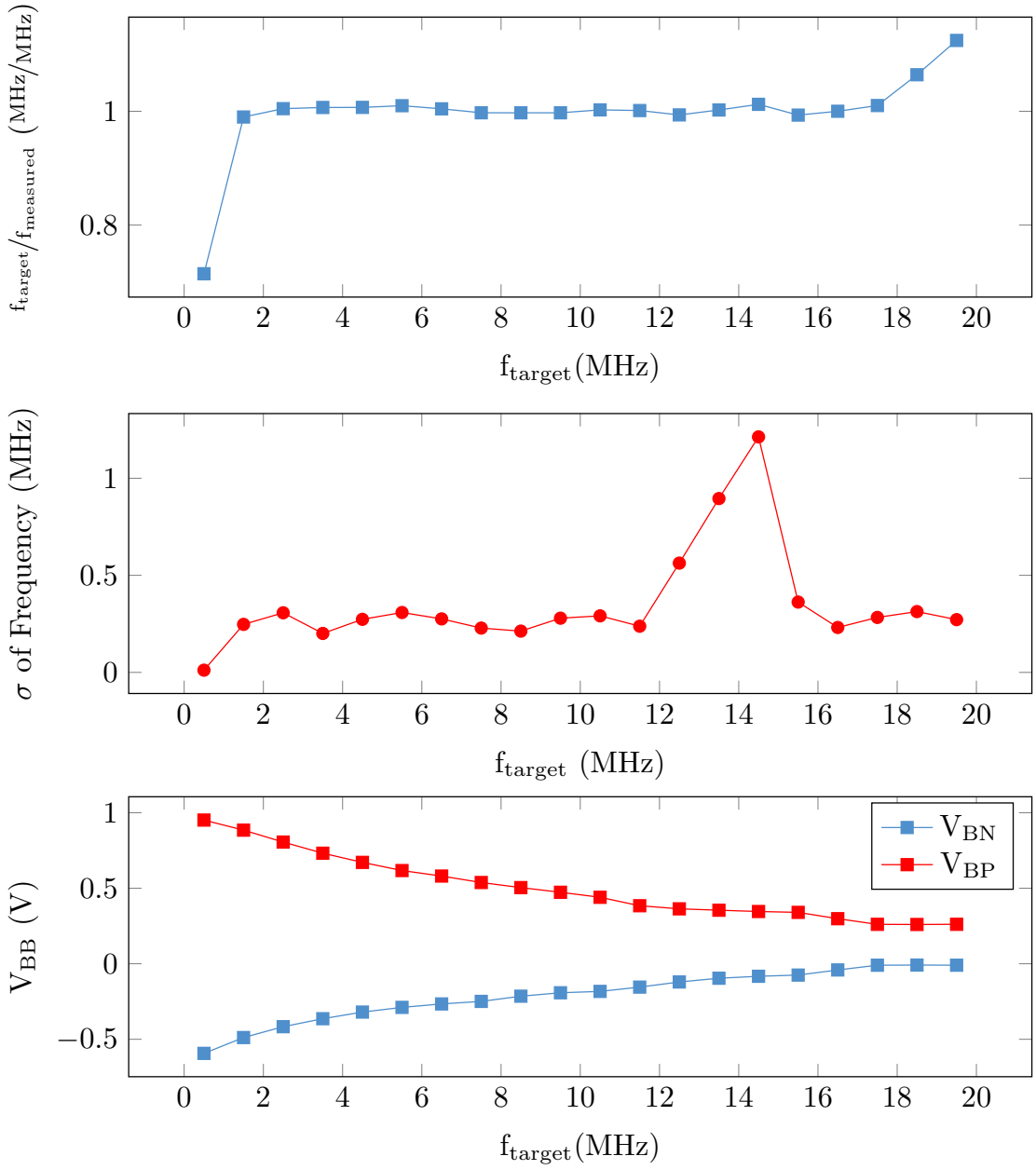



Figure 6.14 – Top: Ratio of target frequency to measured frequency when sweeping the target frequency from 500 kHz to 19.5 MHz at room temperature with a typical die. Center: the corresponding frequency standard deviation. Bottom: the corresponding NMOS and PMOS bias voltages.

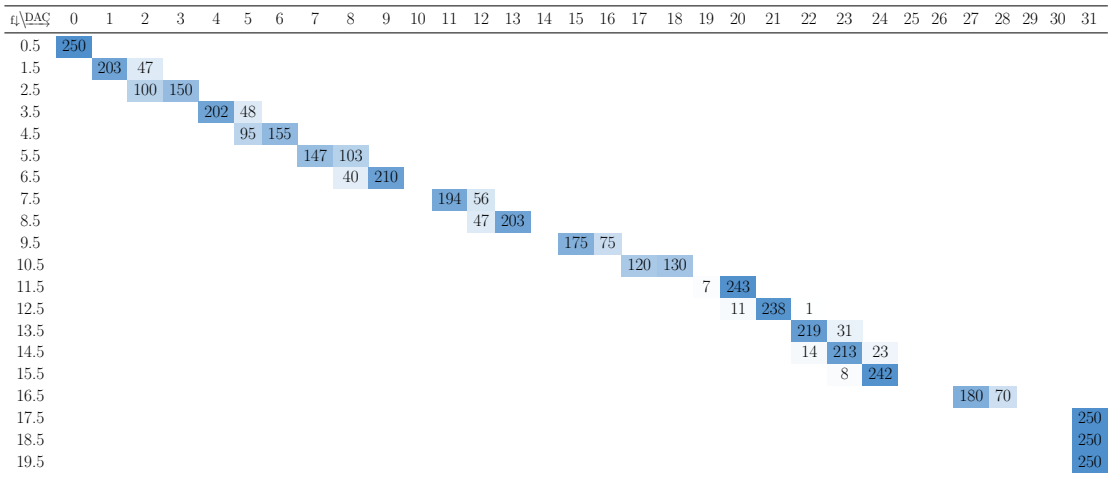



Figure 6.15 – DAC code histogram for the frequency sweep shown in Fig. 6.14, generated by reading back the current DAC code 250 times during regulation.

$$f_{ratio} = \frac{f_{target}}{f_{measured}}$$

The data has been acquired using a digital oscilloscope, measuring frequency averaged over at least 100 samples. As we can see, we achieve a decent regulation performance, with the frequency error within single digit percents, except for the two extreme ends: at these points the requested frequency is too low or high to be achievable through the DAC. Note that we achieve this linearity beside the voltage drop responsible for the discontinuity shown in Fig. 6.13, the FLL is compensating the circuit for the voltage variation as intended.

Below we plot the variation ($\sigma$) for the measured frequency. The observed standard deviation depends on the how close the target frequency requested matches the achievable frequency achievable due to the discrete step size of the current DACs integrated in the bias generator. An extreme case can be found at 14.5 MHz where the requested frequency by chance closely matched a whole DAC code, with the result that the regulation over- and undershoots. This can be observed in Fig. 6.15 which plots the histogram of the DAC codes applied to the bbgen, measured by reading the corresponding register 250 times during regulation. For the 14.5 MHz case we can observe a spread over three DAC codes.

### 6.5.4 System Power

Lets now have a glance on the System power: Table 6.1 shows the circuit currents both at full retention and at full operation running at 8 MHz. The power savings of the retention mode are significant: the total system power drops by a factor of 320x at the nominal 25 °C 0.6 V operating mode when moving the system from a full memory configuration at the 8 MHz operation mode

Table 6.1 – Measurements for Fast and Slow modes of leakage for the Nakayama SoC for the typical condition.

| | TT, 0.6 V, 25 °C<br>Full System |
|---|---|
| Retention | 781 nA |
| Operation (8 MHz) | 250 µA |
| Reduction | 320.1x |

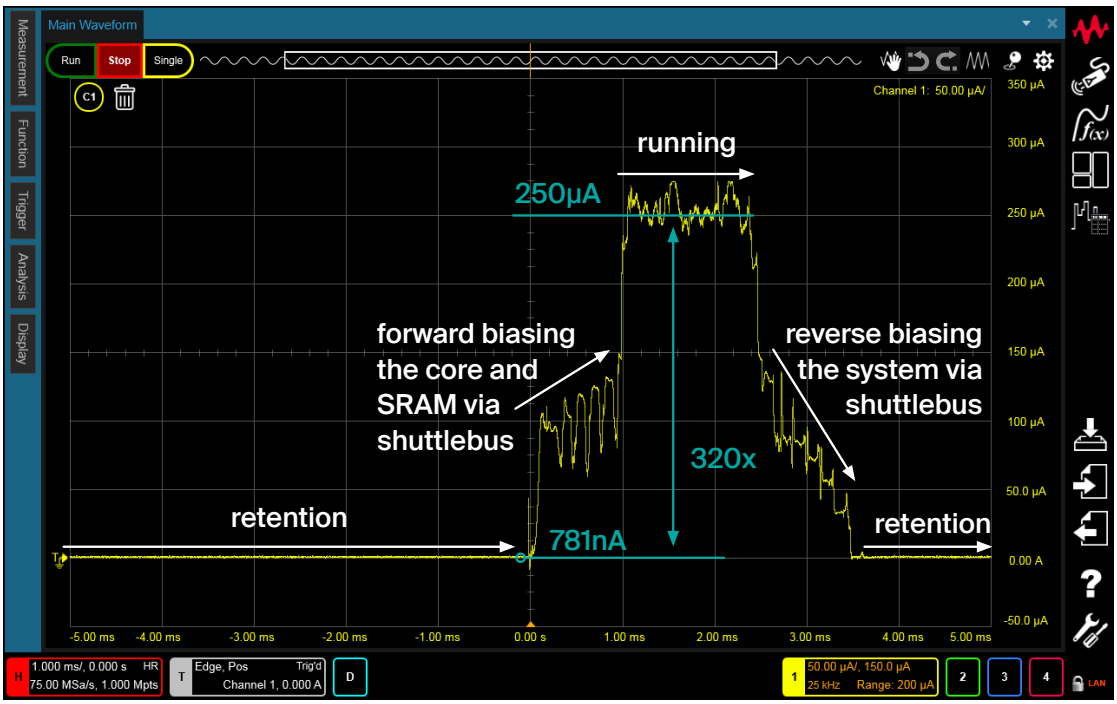



Figure 6.16 – Mode switching on the Nakayama SoC: The system is in retention, enables the SRAM banks, runs for 1.5 ms and returns back to retention.

towards the retention mode.

Figure 6.16 shows the current trace of the mode switch of the circuit for the 25 °C 0.6 V case: The system starts in retention, wakes up over the time span of 1.1 ms, operates at full speed for 1.3 ms while then returning back to retention over the period of 1.1 ms. We can observe the discrete steps of the memory banks being switched on when pushing the system forward and similarly when pushing it back into retention.

## 6.6 Conclusion

In this chapter we have presented the use of a secondary FLL loop in order to provide a target frequency control knob to the user. We generate the system clock with a programmable length

ring oscillator allowing for margin adjustment which is speed controlled by the FLL jointly with the digital system through the body.

We shown that a secondary FLL solves the overcompensation tendencies from the pure on-current based biasing scheme. The System regulates towards an *average* frequency close to the programmed value, however is limited by the coarse steps of 5 bit bias DAC, resulting in larger than necessary variation of the frequency during operation. An improved version could integrate a fast sigma delta modulation scheme of the DAC codes in order to suppress the inherent frequency variation by keeping the error in tighter bounds.

Nakayama system power measurements have shown the effectiveness of the biasing system, reducing the system power to 1/320th when switching from operation to retention. The scheme of banking the memories into biasing banks of different sizes in Nakayama is an easy method to provide the programmer with an interface to exploit the gains of biasing based sleep modes.

# 7 Snaefellsjokull: Worst Case Oriented Body Bias

In this chapter, we propose a new biasing system for a typical sensor node. The integrated microcontroller with body bias support alternates between an active mode with moderate, well defined, processing requirements and a sleep mode in which no processing is required, but data in the memories must be retained.

While Calanda and Nakayama strive to achieve near perfect timing compensation by pinning the on-currents, Snaefellsjokull follows a different biasing approach, with the objective to keep the system as simple as possible. The goal of perfect timing compensation with minimum power across all PVT conditions and modes is replaced in favour of an approach where we focus on achieving the best retention for the worst case. This approach is motivated by a manufacturers perspective which is ultimately only interested in improving the worst case numbers in the data sheet while maximising the yield.

Specifically, we investigate a system that implements a fast operation mode where the circuit is pushed to its nominal bias operating point as well as a reverse bias mode for retention. The bias is in this case not selected to compensate, but rather just such, that the worst case achieves the performance required while accepting worse performance in the other corners. For the operation mode that means that the slowest corner is fast enough to achieve the response time dictated by the algorithm and application. Our main concern for the implementation of the retention mode however is to achieve the lowest leakage performance for the worst, most leaky, corner.

## 7.1 System Architecture

The Snaefellsjokull system is depicted in Fig. 7.1. The actual system is very similar to the one used in Calanda (compare Fig. 5.2), and shares the same icyflex 32 bit RISC core, 64 kB of SRAM, 4 kB of ROM as well as the same standard peripherals.

The difference lies in the biasing concept where the on-current biasing from Section 5.2 is

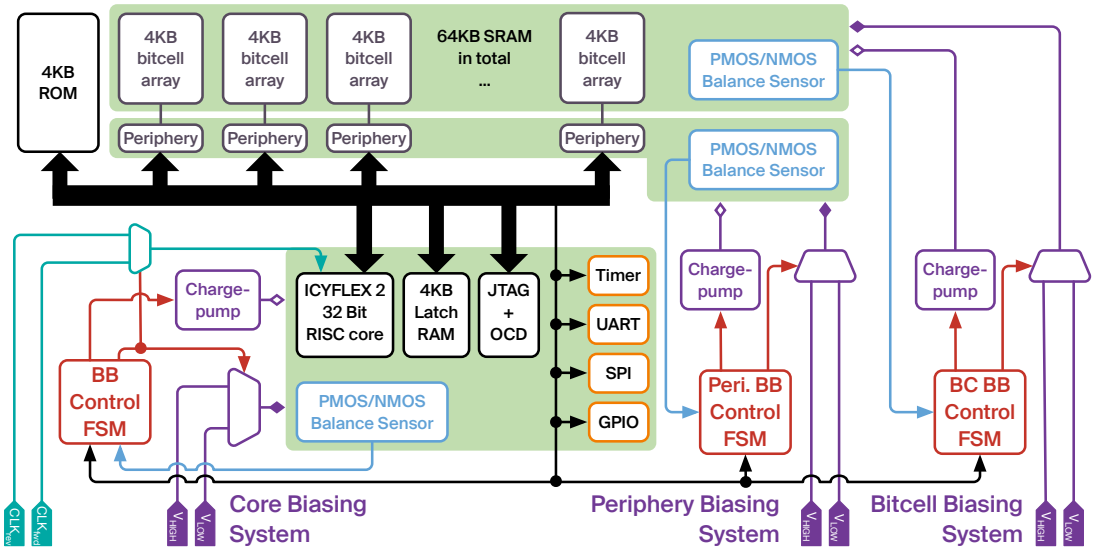



Figure 7.1 – Snaefellsjokull system overview: The system itself is very similar to the Calanda design shown in Fig. 5.2, but the biasing system has been replaced with a simplified one, directly employing a charge pump on the well.

replaced by a simplified system which sets the positive PMOS bias by multiplexing between fixed retention and operation voltages, for the sake of simplicity provided externally. The negative NMOS bias is then set on chip by a charge pump, controlled with the output of a PMOS/NMOS balance sensor. The result is a system capable to reach two predefined operating modes with NMOS bias following the PMOS such that the transistor on-currents are balanced.

## 7.2 Biasing Concept

The requirements for the two operating modes are quite different: for retention we only care about the worst case leakage which should be as low as possible. The system timing in this retention mode is less important: as the majority of the logic is clock gated with only a small wakeup counter remaining active. Nevertheless the bias has to be kept sufficiently balanced to keep the static noise margin in all sequential elements of the core sufficiently large.

On the contrary, for operation we care mostly about speed: in the slowest case we need to be able to achieve the operating frequency specified at design time across all corners. Hence, the bias has to be chosen such that the slowest corner becomes fast enough to achieve the desired frequency without timing violations. Core power is dominated by the dynamic component if the circuit is not pushed too far forward where the leakage can become excessive for some PVT conditions (compare Fig. 3.3).

Besides these functional requirements for the two operating modes, the biasing system should also allow for an easy integration, ideally in an all digital flow. This means that sensors placed

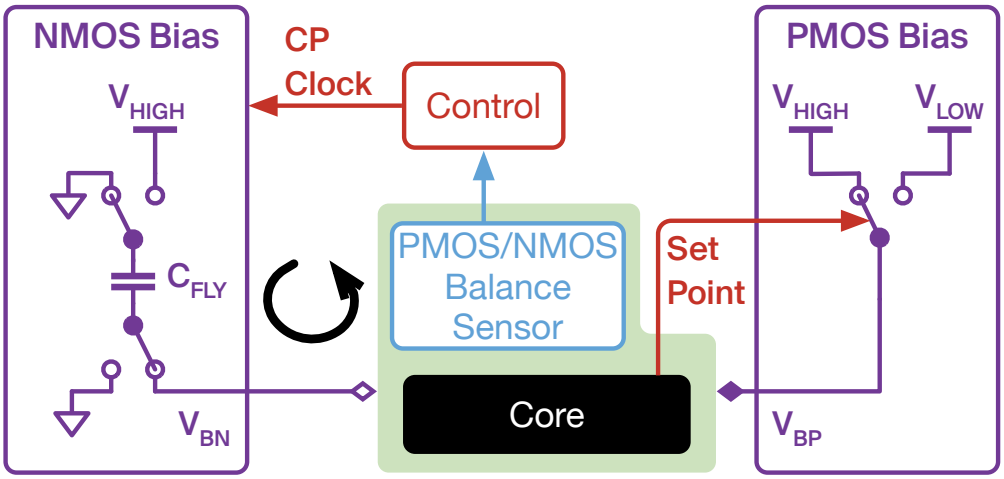



Figure 7.2 – Simplified direct charge pump biasing concept. The target bias is set by selecting the PMOS bias through an analog MUX. The change in bias is sensed by a PMOS/NMOS current balance sensor, enabling/disabling the NMOS charge pump clock.

within a power domain should be standard cell compatible. Further it should be possible to easily size up the functional components to accommodate for different bias domain sizes throughout the design. Finally, size is a concern as the overhead of adding biasing domains to a system should be as low as possible in order to allow for a fine grained power management through the application of body biasing (as for example the banked memories of Nakayama).

The result of these requirements is the biasing system depicted in Fig. 7.2. We use the most simple implementation of a PMOS biasing system possible, an analog MUX which switches between the forward and reverse bias voltage $V_{Low}$ and $V_{High}$, respectively. For our proof-of concept design these are supplied externally, but since the well leakage currents are low, two LDOs could be used with a very limited penalty on the overall system. For supplying the NMOS bias we need a charge pump to generate the negative voltage. We propose to use the duty cycle of the pump to implement the regulation of the NMOS bias such that we match the drive strength to the PMOS, based on the output of a PMOS/NMOS current balance sensor placed inside the bias domain.

In the most simple implementation this current balance sensor can directly enable and disable the charge pump clock, based on the sensor output. When changing the PMOS bias voltage through the switch the trigger point of the sensor shifts, resulting in different charge pump duty cycles.

### 7.2.1 $I_{ON}$ and $I_{OFF}$ Considerations in Operation and Retention

During retention, the system power is dominated by the leakage current $I_{OFF}$ as the majority of the system is clock gated with only a minimal subsystem running, implementing the ability to

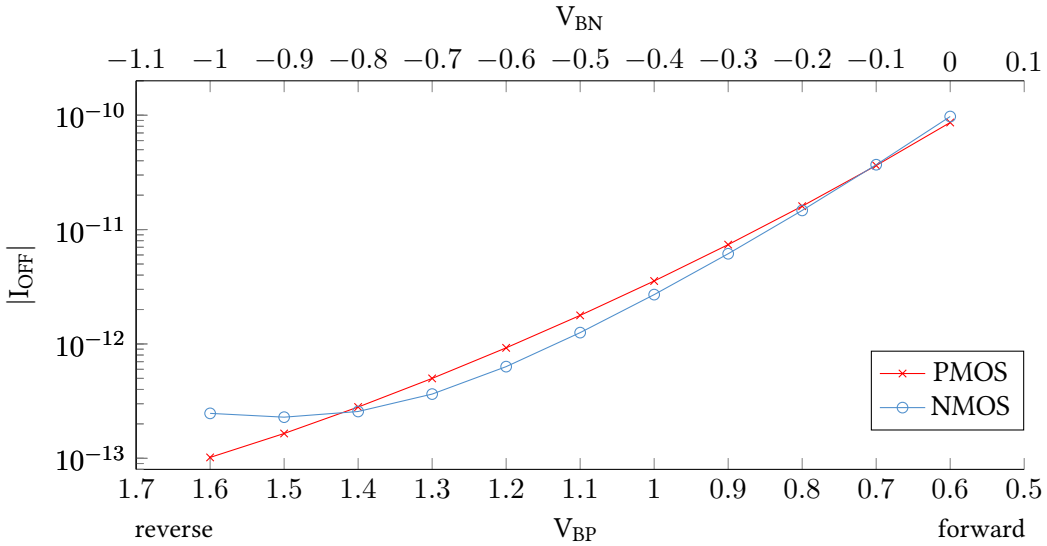



Figure 7.3 – Off currents through a single minimum sized NMOS/PMOS transistor when sweeping the body voltage.

wake up based on a timer or interrupts from a potential sensor subsystem. Timing is less critical as the wakeup circuitry can be implemented with significant setup margin, and therefore also degradation of the balancing between the PMOS and NMOS is acceptable in extreme PVT cases. The off-current $I_{OFF}$ is shown in Fig. 7.3 for PMOS and NMOS, respectively. For far reverse bias the NMOS manifests a flat curve, resulting in a near constant leakage behaviour over 300 mV in the bias voltage range of −0.8 V down to −1 V with the minimum close to −0.9 V. In the context of the charge pump based retention bias system this allows for a fairly wide hysteresis of the charge pump operation duty cycle without a penalty in system leakage.

During operation, the main concern is the system timing, the critical paths have to at least achieve the speed set by the system clock. A balanced NMOS and PMOS $I_{ON}$ is desirable in order to reach balanced rise and fall timing so that neither NMOS driven pull-down networks nor the PMOS driven pull-up networks dominate the overall system timing. The corresponding $I_{ON}$ currents for a minimum sized transistor are shown in Fig. 7.3 when shifting the bias. We observe the expected offset between NMOS and PMOS due to the higher carrier mobility. In our proof of concept design we use the same standard cell library as for Calanda and Nakayama which is designed to utilise the biasing also for compensation of the mobility differences, resulting in the need of a reverse bias of approximately 300 mV for the NMOS relative to the PMOS bias.

### 7.2.2 Well Leakage Considerations

The well insulation is formed by reverse biasing the built in substrate diode of the p-n junction on the contact points of the PWELL and the NWELL. The leakage currents through this diode are the main contributor for the well leakage and hence set the current driving capability

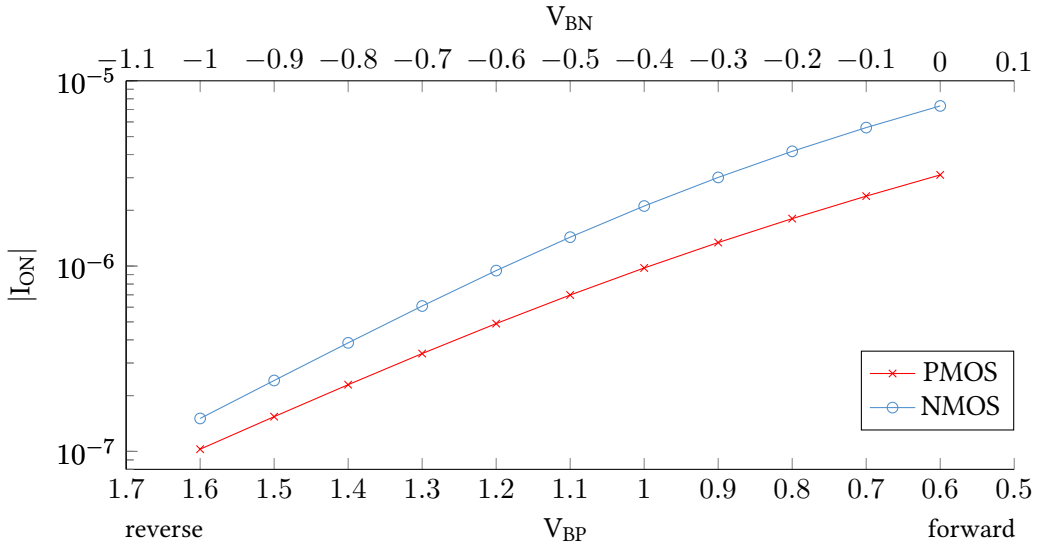



Figure 7.4 – On currents for a single minimum sized NMOS/PMOS transistor when sweeping the body voltage.

requirement for the charge pump. The diode size, and hence the leakage, scale with the size of the bias domain since it is formed by the abutment of the NWELLs inside both the standard cells as well as the filler cells. However, when increasing the size of the bias domain we also scale up the size of the capacitor formed by this reverse biased diode. As a result the well decay time can be considered constant for a given PVT corner as well as area independent.

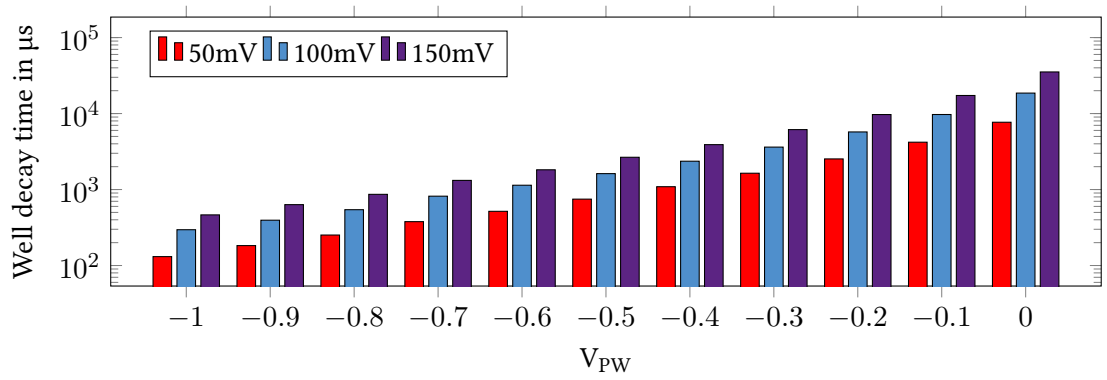
Figure 7.5 shows the simulated PWELL decay times for changes in the well voltage of 50 mV, 100 mV and 150 mV, resulting in a duty cycle for the charge pump in the range of about a millisecond for the typical case and a few microseconds in the worst case.

The leakage and decay time directly determine the sizing of the charge pump and the sampling/regulation time constant, respectively. The charge pump has to be capable to drive the load against the leakage and the regulation has to be sufficiently fast to keep the decay between two sensing events in an acceptable range. We are interested in keeping the sensor sample rate as low as possible to reduce the power impact as much as possible.

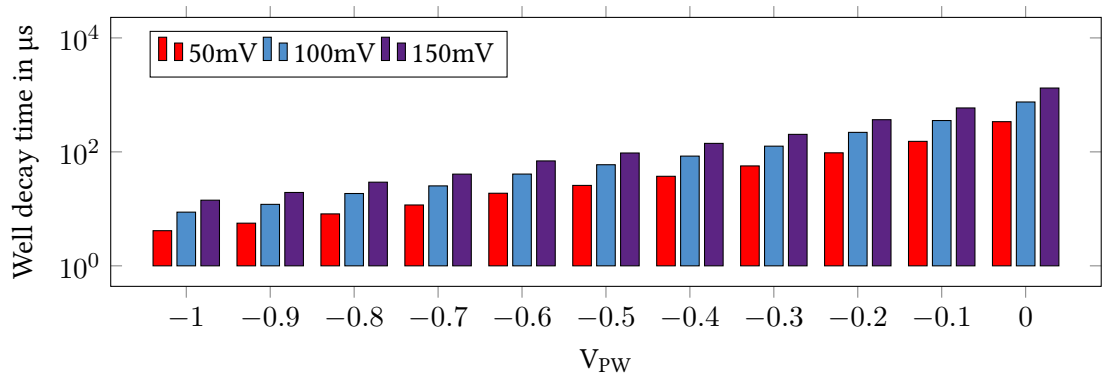
Based on the worst case time constant shown in Fig. 7.5, combined with the 400 mV bias voltage range with near constant NMOS leakage from Fig. 7.3 we can just get away with sampling the well bias during retention with a 32.768 kHz clock, without significant leakage degradation due to the resulting bias ripple.

### 7.2.3 Sensing

In order to control the charge pump, some kind of sensing mechanism is needed. For operation the choice is fairly straight forward: similarly to Calanda and Nakayama a balanced drive




(a) Typical case - TT / 25°C




(b) Worst case - FF / 85°C

Figure 7.5 – Simulated p-well decay times, starting with an initial condition of $V_{PW}$ and allowing for a decay of 50 mV, 100 mV, and 150 mV.

current for NMOS and PMOS is desirable in order to obtain balanced rise and fall times to achieve a well controlled timing behaviour of the digital circuit across PVT. Hence, a sensor detecting the balance between NMOS and PMOS on currents is a good fit.

For the leakage dominated reverse bias case it might seem ideal to try to sense the off currents $I_{OFF}$ directly or through some kind of proxy placed in the bias domains. However this is not trivial to do, due to the extremely low magnitude of these currents. In particular, when considering the duty cycle requirement from the well leakage analysis the corresponding time constraint on the sensor becomes challenging to meet in deep reverse sub-threshold operation. Further, even a small amount of variation between two devices results in substantial variation of the off currents due to the exponential nature of the sub threshold operation. Instead we propose to use $I_{ON}$ as a proxy for $I_{OFF}$ even during retention. This has several benefits: first, the on currents are significantly larger than the off currents and as such much easier—and faster—to measure. Second, contrary to the $I_{OFF}$ shown previously in Fig. 7.3 the $I_{ON}$ is strictly monotonically increasing with a more forward bias, allowing for a simple and stable regulation loop. Finally, this approach allows to reuse the same sensor for both operating modes, significantly simplifying the biasing system and its integration which is the main objective of the system.

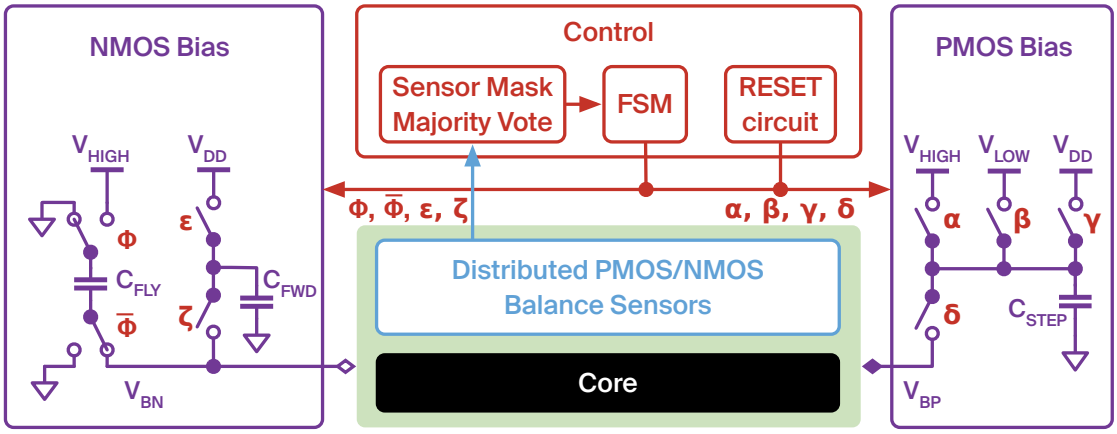



Figure 7.6 – Snaefellsjokull biasing system

## 7.3 Bias System Implementation Details

In this section we discuss the implementation details of the components of the Snaefellsjokull biasing system, as depicted in Fig. 7.6. The figure denotes control signals generated by the FSM with greek letters. The negative charge pump for the NMOS is controlled by the clock $\phi$ and its inversion $\overline{\phi}$ while the NMOS forward switch uses the control signals $\epsilon$ and $\zeta$. The PMOS switch is controlled with $\alpha$, $\beta$, $\gamma$, and $\delta$.

To improve the robustness against on-chip variations, the implemented biasing system uses multiple sensors, distributed over the whole bias domain. The binary sensor outputs are combined by adding them up and comparing the result against a user configurable threshold. Further, a masking capability allows to remove non-functional sensors. The bias switches include capacitors to smoothen the transition and additional switches to actively speed up the transitions between all modes.

### 7.3.1 Distributed NMOS/PMOS Balance Sensor

We propose a sensor based on the simple topology depicted in Fig. 7.7, which detects imbalances between the NMOS and PMOS drive current. The difference in drive strength is directly reflected in the potential $V_{center}$ seen on the center node. This node is directly connected to the input of an inverter exploiting the high gain of CMOS to achieve a very well defined triggering point very close to the equilibrium of the two drive strengths. This circuit is sufficient to implement a very simple control loop as $V_{out}$ can be directly used as the enable signal for the charge pump.

Figure 7.8 shows a Monte Carlo simulation of this concept in the unbalanced fast/slow and slow/fast corner relative to the typical case. We observe the shift expected to rebalance the devices: In the SF corner the fast NMOS requires a more reverse bias while the slow NMOS of the FS sees a more forward bias respectively. As expected, local variations covered by the Monte Carlo simulation result in a gaussian distribution around the center point, suggesting

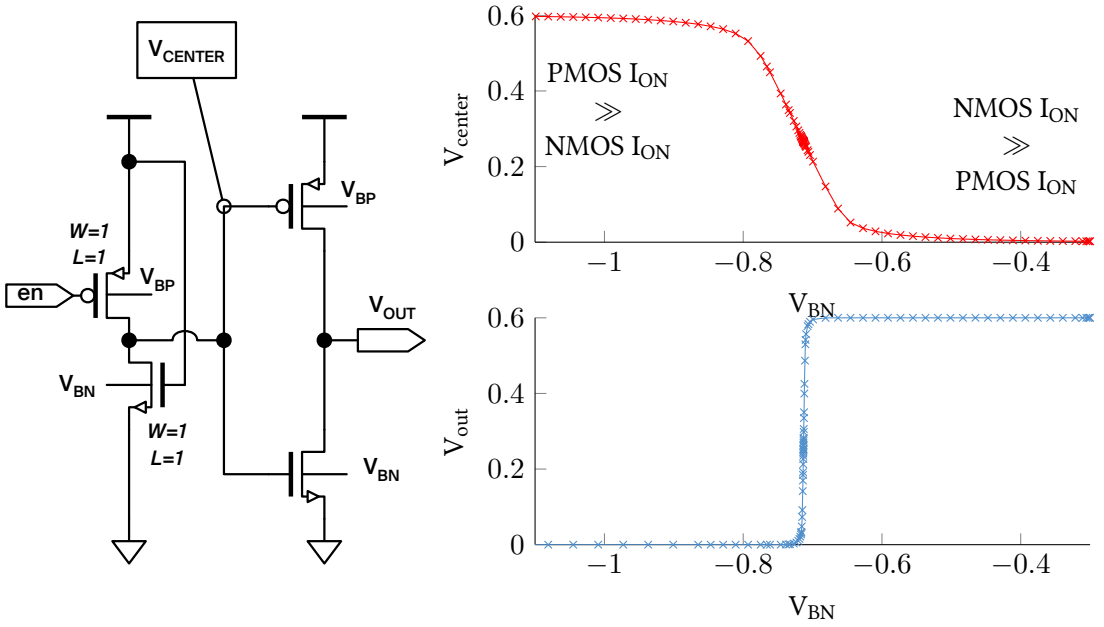



Figure 7.7 – Sensor tripping point for when sweeping the NMOS bias while keeping the PMOS constantly reverse biased by 300 mV.

the use of multiple sensors with a threshold or majority vote mechanism in order to achieve a reliable on current bias tracking.

The obvious drawback of the simple sensor design in Fig. 7.7is the inherent power consumption due to the NMOS/PMOS pair being operated in short circuit during readout. To reduce the power consumption we propose to integrate a latch within the sensor in combination with a pulse generator that provides a pulse just long enough to reliably sample the on current ratio. The schematic and layout of the cell, integrating the sensor and a latch, is depicted in Fig. 7.10. With a size of $6.84\,\mu\text{m} \times 1.62\,\mu\text{m}$ the sensor area is comparable with the area of a flip-flop.

The potential $V_{center}$ is sampled on the latch input where the internal state trips around the point where the drive strength of NMOS and PMOS is approximately identical. We use the pulse generator in Fig. 7.9, to create a pulse of the length needed for the clock signal traversing 20 inverters. The pulse generator is intended to be placed in the same bias domain as the sensors, hence the pulse length scales automatically with the bias condition seen in the domain.

As the tripping point is sensitive to a) local variations and b) noise we propose the distribution of several sensors across the bias domain and combine their binary outputs either with a majority vote or a comparison of the sum against a configureable threshold. Finally, the sensor has been extended with three additional programmable parallel PMOS and NMOS transistors that share $V_{center}$, allowing to either shift the tripping point or increase the sense currents in situ.

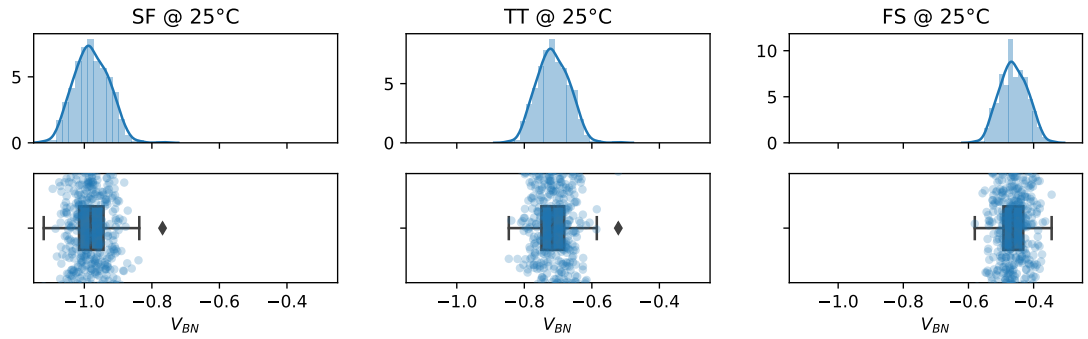



Figure 7.8 – Monte-Carlo Simulation of the sensor trip point voltage for FS, SF and TT corner with the PMOS reverse biased by 300 mV. Top: probability distribution. Bottom: boxplot with the actual simulation results, distributed randomly across y for better visualisation.

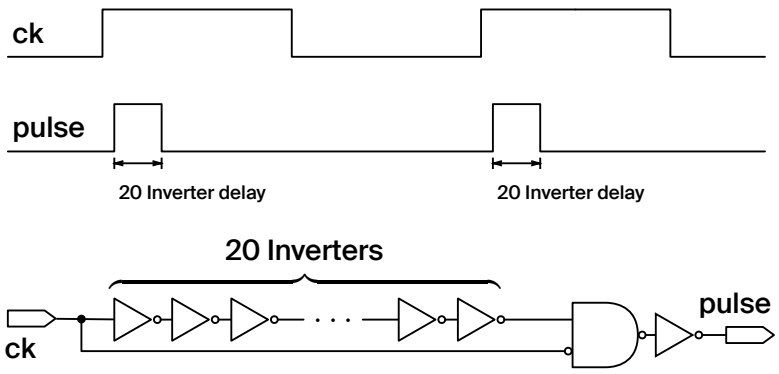



Figure 7.9 – Pulse generator

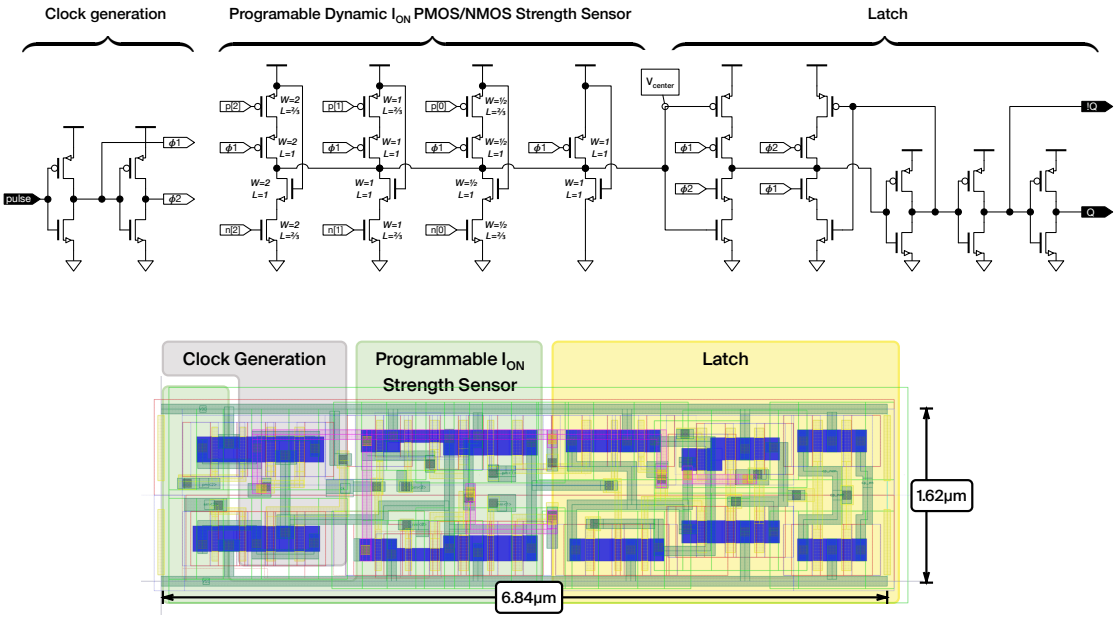



Figure 7.10 – Drive Strength sensor, latching the center voltage of two shorted PMOS / NMOS during a short period of time.

### 7.3.2 PMOS Bias Switch

The bias of the PMOS is controlled using the simple pass gate switch depicted in Fig. 7.11, that allows to select between a retention supply ($V_{High}$), an operation supply ($V_{Low}$) as well as $V_{DD}$ with the latter used during system reset. In addition an output switch and a capacitor are integrated, to smoothly transition between well values by dumping only the charges saved in the cap onto the well. The switch is laid out as a hard macro with all control signals and power rails placed parallel on the top of the macro touching the edges of the macro. This layout configuration allows to easily scale up the drive strength by abutting more instances, adding an area of $112\,\mu m \times 25\,\mu m$ for each slice capable to drive the area equivalent of 100k cells. The signal generation block area is $11.5\,\mu m \times 14\,\mu m$.

The pass gates are controlled by a simple signal generation circuit, that integrates level shifters from $V_{DD}$ to the highest bias voltage in the switch as well as buffered and inverted versions of the input signal. The signal generation circuit has been implemented as a separate layout, designed to abut to to the switch slice.

### 7.3.3 NMOS Bias Charge Pump

Figure 7.12 shows the charge pump layout and schematic, following the architecture previosly implemented in [50]. The design again has been built to allow for abutment to scale the drive strength up as necessary for a given bias domain. For this purpose each slice contains a replica of the full H-bridge configuration used to inverse the polarity of the capacitor $C_{Fly}$. Each slice

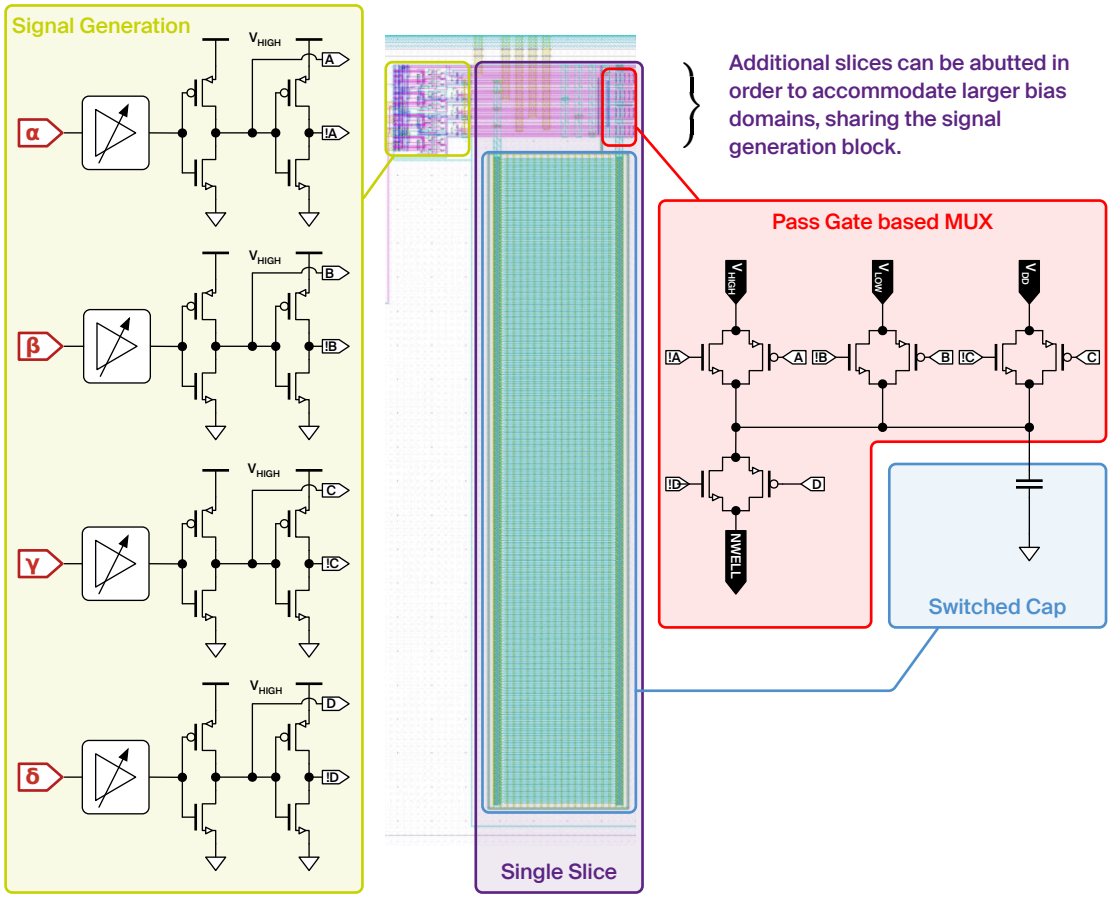



Figure 7.11 – Schematic and layout of the NWELL switch, muxing between the dedicated retention voltage, the dedication operation voltage, and VDD.

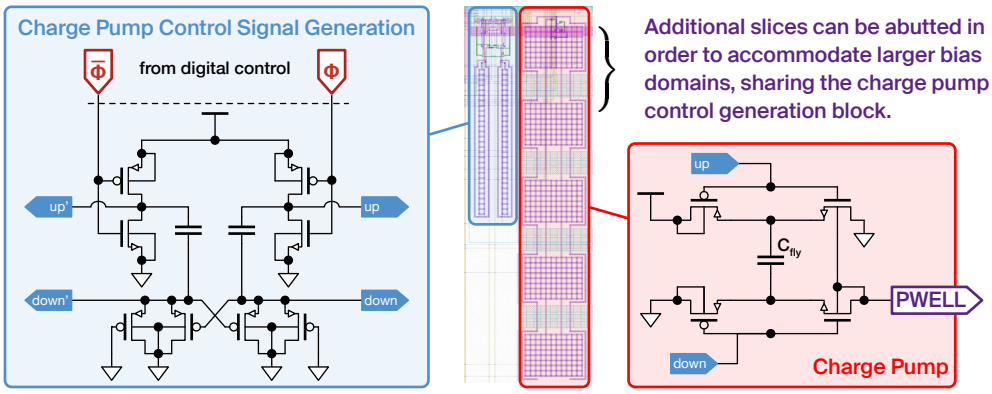



Figure 7.12 – Schematic and layout of the charge pump as well as the corresponding control signal generation circuit.

of the charge pump measures $162\,\mu m \times 32.2\,\mu m$

The control signals are generated by a signal generation macro, which uses a cross coupled bootstrap circuit for the generation of the negative control signal needed on the bottom end of the H-bridge. The circuit is symmetric and can be abutted to the charge pump slices in order to achieve the required connectivity. If charge pump slices are abutted left and right of the macro a dual phase charge pump topology can be constructed. The size of the macro is $100\,\mu m \times 21.4\,\mu m$.

### 7.3.4 NMOS Forward Switch

In principle the charge pump is sufficient to implement a fully operational biasing system due to the well leakage counteracting its operation. However, a dedicated forward switch is desirable in order to allow for a faster and better controlled forward regulation. Figure 7.13 depicts the simple PMOS based forward switch schematic as well as the layout with an area of $100\,\mu m \times 16\,\mu m$. Again, a capacitor is integrated in order to allow for a regulation in discrete steps.

In parallel a secondary reset control is integrated for regulating the well during reset towards ground potential by either enabling the charge pump or pulling the well towards $V_{DD}$ using using PMOS switches. This topology has been used instead of an NMOS based switch directly towards ground to guarantee a high impedance during reverse bias operation when the PWELL voltage is negative.

The required control signals are generated by the differential amplifier highlighted in orange in Fig. 7.13. The amplifier compares the current through a $V_{SS}$ biased NMOS against the current through an NMOS biased by the PWELL voltage of the bias domain. This approach of shifting the balance of the differential amplifier through the bias, has two benefits: first, the comparator can be implemented with just $V_{DD}$ and $V_{SS}$, not needing a negative supply, and second, does

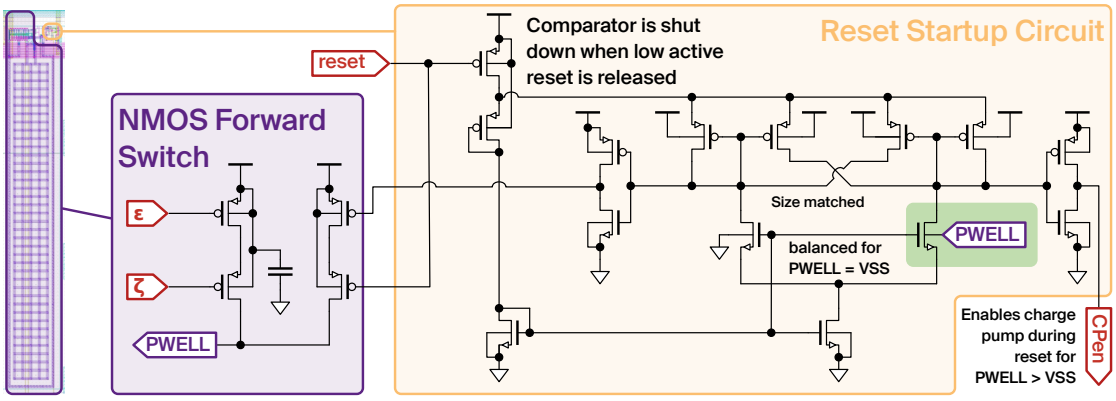



Figure 7.13 – Forward switch and reset bias subsystem.

not need high voltage, thick oxide devices which keeps the control circuit compact.

During reset the output of the amplifier is either used to pull the well towards the supply voltage through the reset branch in Fig. 7.13 or, in reverse direction, by enabling the charge pump clock. When reset is removed the comparator is power gated.

### 7.3.5 Proposed Regulation Loop

In a very simple implementation the PMOS bias voltage could be directly switched between the forward / reverse voltage. In that case the output of the balance sensor could directly control the charge pump for reverse regulation while we could rely on the well leakage to shift back towards more forward bias. However, these switches would be fairly harsh and would momentarily cause significant PMOS/NMOS imbalances reducing the SNR of all sequentials in the design.

Hence, a more controlled regulation is desirable, switching the system in discrete steps: Figure 7.14 shows a simplified FSM for the control switching between operation and retention through a transition state.

#### Operation

During operation the control loop runs with the full system clock. In order to save power the a programmable idle timer is employed, allowing to adjust the sensor readout frequency. Once the sensor trips the charge pump is enabled until the sensor trips back. When the system software wants to send the core into retention a switch is triggered by writing into the corresponding register.

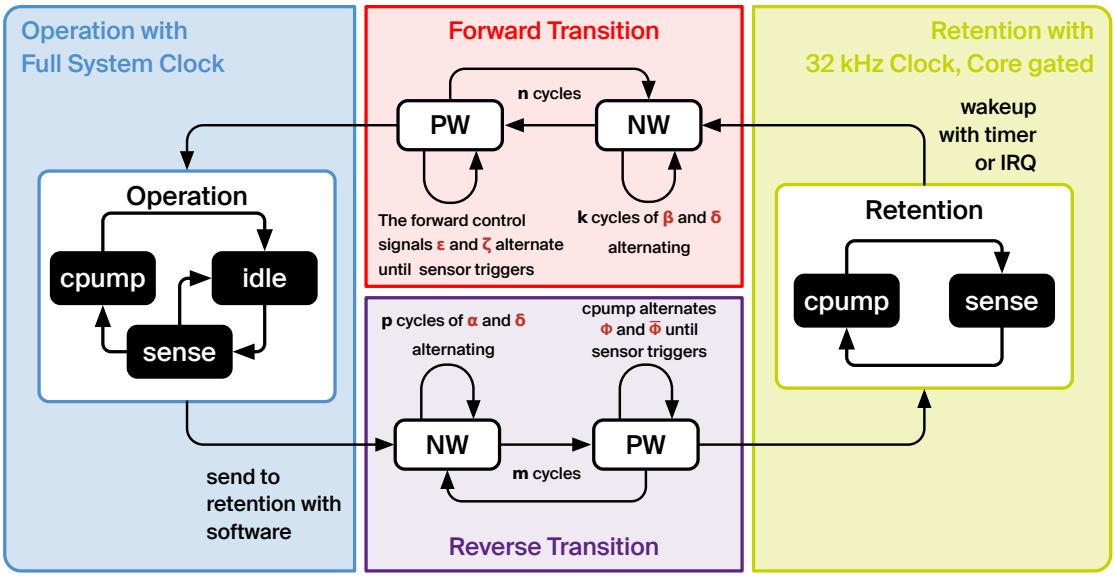



Figure 7.14 – Simplified State Machine of the snaefellsjokull bias control system. The role of the control signals $\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, $\zeta$, $\phi$ and $\overline{\phi}$ is shown in Fig. 7.6. The values of **k**, **p**, **m** and **n** are programmable for in situ tuning.

### Reverse Transition from Operation to Retention

The transition phase is entered in the *reverse NW* state and the PMOS bias switch presented in Section 7.3.2 kicks in: the direct supply from the operation voltage $V_{LOW}$ is disconnected and the capacitor is charged with the retention voltage $V_{HIGH}$ for **p** cycles during which the control signals $\alpha$ and $\delta$ alternate. We then transfer control to the *reverse PW* state which operates the NMOS charge pump from Section 7.3.3 until the sensor triggers to switch back to the *reverse NW* state to shift the NWELL further towards $V_{HIGH}$. This process repeats **m** times with **m** being programmable through a register. The user has to choose the value large enough so that we can be sure to have reached $V_{HIGH}$ on the well. In that case $V_{HIGH}$ is directly connected by enabling the control signals $\alpha$ and $\delta$ at the same time. We enter retention, switching the regulation clock to the slow 32.768 kHz clock running a sensing every cycle and enabling the charge pump if needed to maintain the PWELL bias.

### Retention

During retention we enter the *sense* stete where the sensors are continuously sampled with the slow 32.768 kHz clock until the sensor reaches the threshold. As soon as the sensor triggers the *cpump* state is entered and the pump runs for as many cycles as needed to toggle the sensor back.

#### Forward Transition from Retention to Operation

Wakeup is triggered through a timer or by an external interrupt signal at which point we enter the transition procedure in the opposite direction. We enter the *forward NW* state where $V_{HIGH}$ is disconnected from the NWELL by the PMOS bias switch from Section 7.3.2 and operated in the opposite direction, pulling the NWELL **k** times towards $V_{LOW}$ by alternating between $\beta$ and $\delta$. Afterwards we switch to the *forward PW* state where the control signals $\epsilon$ and $\zeta$ alternate, slowly pulling the PWELL towards ground until the sensor triggers again. This is repeated **n** times. Finally $V_{LOW}$ is connected directly to the well by enabling the control signals $\beta$ and $\delta$ at the same time. We are back to active operation, with the regulation running at full system speed.

## 7.4 Bias System Operation

Figure 7.15 shows a transient simulation of the Snaefellsjokull biasing system with the PMOS bias set in 100 mV steps from 0.8 V to 1.3 V in the typical corner at 25 °C. We start with the pwell at ground potential with the sensor immediately enabling the charge pump due to the imbalance in bias. The charge pump is operating continuously until a balanced on-current is reached. After settling the average NMOS bias voltage follows the expected 100 mV shift expected by the 100 mV PMOS bias step size. The remaining ripple stays below 100 mV, indicating a good match of the charge pump drive strength to the circuit size.

We can further observe that the duty cycle decreases for the reverse bias, which is expected, considering the increasing well leakage presented in Fig. 7.5. Consequently, this results in an increase of dynamic power dissipated in the charge pump for more reverse cases. Even worse, the charge pump conversion efficiency directly depends on the voltage step when (dis)charging the capacitor [81, p. 23ff.] which decreases further the more reverse we go, further degrading the efficiency of the circuit.

## 7.5 System Retention Power

In the following we analyze the system retention power, based on a digital domain of 100k NAND gates which for we assume to be clock gated for retention, i.e., the dynamic power component of the digital circuit is omitted. All other biasing system related parameters are considered, i.e., we includes the NWELL leakage, the digital system leakage, the biasing system overhead for the sensors, and the charge pump.

Figure 7.16 shows the accumulated system retention power for a system of 100k NAND-Gate equivalents across the TT, FF, FS, SF, and SS corners for −40 °C, 25 °C, −40 °C while sweeping the bias voltage. As expected the high temperature case dominates the overall power, with the FF corner performing the worst, consuming 27.1 µW without biasing which can be reduced by

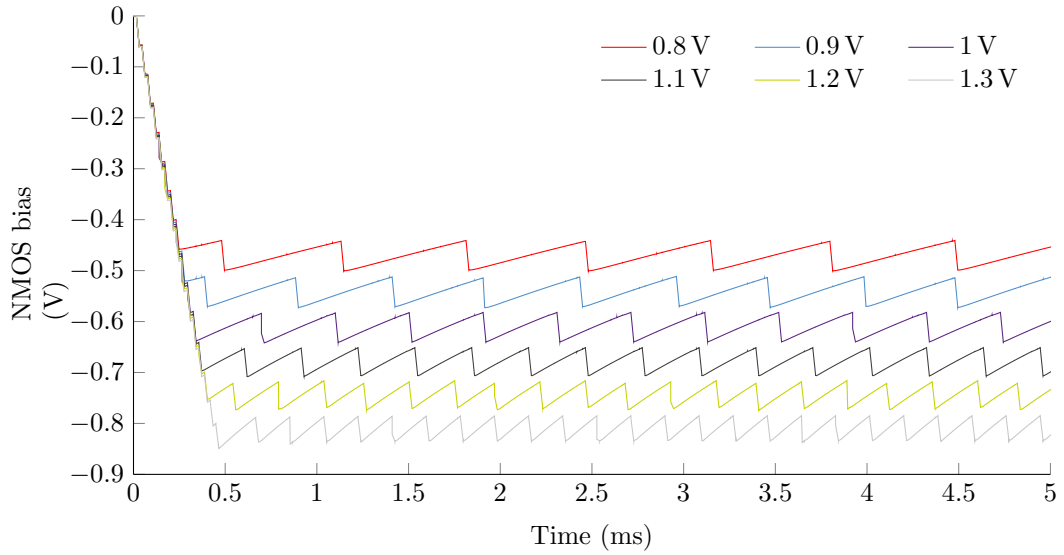



Figure 7.15 – Transient simulation of the Snaefellsjokull control loop of the PMOS/NMOS balance sensor together with the charge pump for a PMOS bias from 0.8V to 1.3V, plotting the corresponding NMOS bias generated by the charge pump.

a factor of ten to 2.7 µW applying a 600 mV reverse bias. At room temperature we observe a minimum power point somewhere around a bias of −300 mV. At −40 °C the behaviour inverts with the power consumption increasing the more reverse we go, except for the FF corner.

The reasons for this become apparent when analysing the contributors as shown in Fig. 7.17. The power data is split into the optimisation target, i.e. the digital leakage, the PMOS well leakage, the balance sensor and the associated pulse generation, the charge pump control, and the charge pump which is sub-summing both the overhead in generating the negative voltage as well as the leakage power of the NMOS well.

At −40 °C the digital leakage is so low that the overhead of generating the bias voltage for all corners, except FF outweighs the benefits. However, with a consumption in the nW-range this overhead is likely of least concern for most applications, even with the observed increasing power consumption for more reverse operation due to a shorter duty cycle of the pump and the decreasing voltage step.

If we consider the room temperature case of 25 °C we observe that the digital leakage becomes significant enough that the utilisation of the biasing system becomes beneficial for retention across all corners, with a minimum power point manifesting at a reverse bias of approximately −300 mV while further reversing results in diminishing returns with again the biasing system overhead dominating.

Finally, for the high temperature case of 85 °C we observe the benefits of full reverse operation: all contributors but digital leakage and the charge pump consumption can be neglected and full

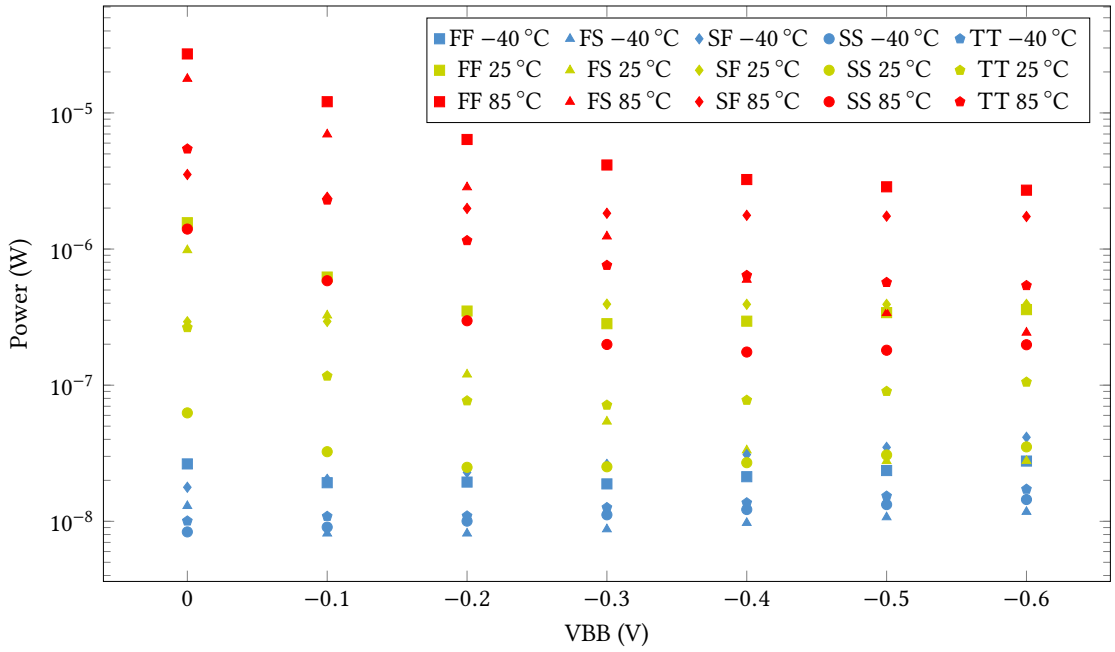



Figure 7.16 – Accumulated retention power for the whole system, including the leakage power for the equivalent of 100k gates, the power of the charge pump, the sensor circuitry as well as the well NWELL leakage.

reverse operation at −600 mV results in the lowest power consumption for all corners except the SS case.

Table 7.1 presents the leakage power components, including the biasing system overhead as well as the well leakage when sweeping the bias $V_{BB}$ reversely from 0 V down to −0.6 V by applying $V_{DD} - V_{BB}$ to the NWELL. For each process and temperature corner the minimum power bias is marked and placed in the *$V_{BB}$ adaptive* column. Considering the switched single voltage approach the manufacturer or user may now face multiple potential optimisation strategies

### Optimize for the worst worst case

In this case our intention is to find one bias value which is pushing the overall worst case across all corners down as far as possible - this approach could be used if the intension is to decide on the retention PMOS bias voltage at design time. The total system leakage power is dominated by the fast corner at 85 °C and benefits from a far reversed bias marked with ■ in the table, resulting in a power reduction by a factor of 10. The column *WC opt.* compares the effectiveness of this approach relative to the ideal adaptive one, showing that this approach is near optimal for one third of the cases while slightly degrading the rest, with an extreme of a factor of 2.32 for the SF case at −40 °C which is likely acceptable due to the anyway very low leakage realised at low temperature.

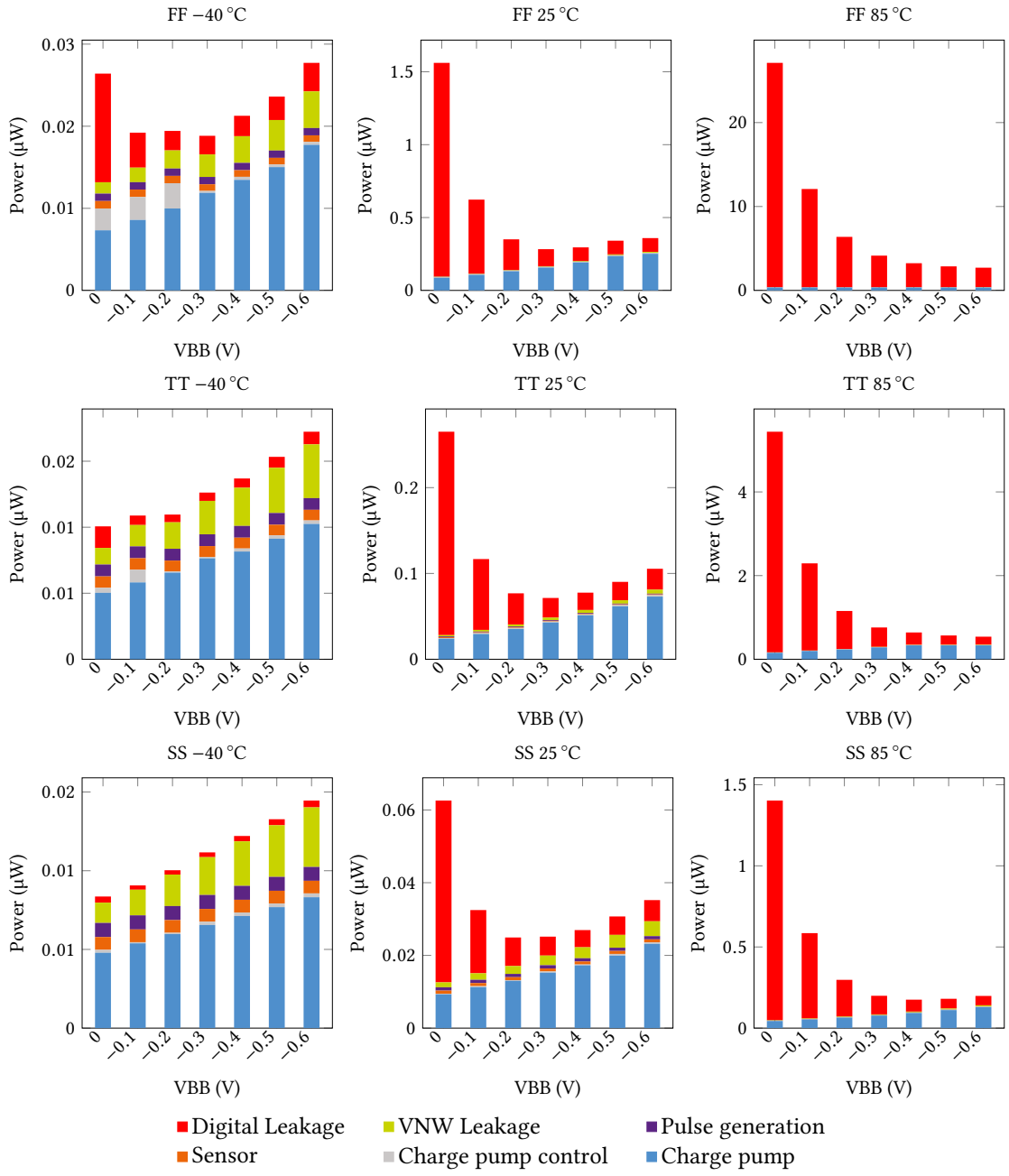



Figure 7.17 – Breakdown of the power contributors across temperature for the equivalent of 100k gates across PVT.

Table 7.1 – Leakage Power when reversing the bias

| P&T Corner | | Full System Power (µW) at $V_{BB}$ 0 V | −0.1 V | −0.2 V | −0.3 V | −0.4 V | −0.5 V | −0.6 V | adaptive |
|---|---|---|---|---|---|---|---|---|---|
| −40 °C | SS | 0.0084 | 0.0091 | 0.0100 | 0.0112 | 0.0122 | 0.0133 | 0.0145 | 0.0084 |
| | FS | 0.0130 | 0.0081 | 0.0081 | 0.0087 | 0.0097 | 0.0107 | 0.0117 | 0.0081 |
| | TT | 0.0101 | 0.0109 | 0.0110 | 0.0126 | 0.0137 | 0.0153 | 0.0172 | 0.0101 |
| | SF | 0.0178 | 0.0203 | 0.0229 | 0.0262 | 0.0309 | 0.0349 | 0.0414 | 0.0178 |
| | FF | 0.0264 | 0.0192 | 0.0194 | 0.0188 | 0.0213 | 0.0236 | 0.0277 | 0.0188 |
| 25 °C | SS | 0.0626 | 0.0325 | 0.0249 | 0.0252 | 0.0270 | 0.0307 | 0.0352 | 0.0249 |
| | FS | 0.9820 | 0.3255 | 0.1199 | 0.0541 | 0.0332 | 0.0277 | 0.0278 | 0.0277 |
| | TT | 0.2651 | 0.1168 | 0.0768 | 0.0714 | 0.0776 | 0.0902 | 0.1054 | 0.0714 |
| | SF | 0.2925 | 0.2940 | 0.3340 | 0.3939 | 0.3928 | 0.3925 | 0.3925 | 0.2925 |
| | FF | 1.5612 | 0.6229 | 0.3505 | 0.2830 | 0.2953 | 0.3412 | 0.3590 | 0.2830 |
| 85 °C | SS | 1.4022 | 0.5852 | 0.2974 | 0.1992 | 0.1752 | 0.1806 | 0.1982 | 0.1752 |
| | FS | 17.8217 | 6.9550 | 2.8446 | 1.2384 | 0.5957 | 0.3390 | 0.2433 | 0.2433 |
| | TT | 5.4395 | 2.2927 | 1.1545 | 0.7598 | 0.6407 | 0.5689 | 0.5396 | 0.5396 |
| | SF | 3.5293 | 2.4000 | 1.9881 | 1.8314 | 1.7694 | 1.7444 | 1.7338 | 1.7338 |
| | FF | 27.1162 | 12.0851 | 6.3808 | 4.1449 | 3.2412 | 2.8655 | 2.7081 | 2.7081 |
| MAX | | 27.1162 | 12.0851 | 6.3808 | 4.1449 | 3.2412 | 2.8655 | 2.7081 | 2.7081 |
| MIN | | 0.0084 | 0.0081 | 0.0081 | 0.0087 | 0.0097 | 0.0107 | 0.0117 | 0.0081 |

### Optimize for the worst case across PVT

In this scenario the manufacturer would determine the corner at production time and trim the retention voltage accordingly. Again, looking into Tab. 7.1 we observe a far reverse bias for all process corners but SS as shown in Tab. 7.3.

These are fairly small gains in the least leaky corner while adding a significant amount of testing cost during manufacturing.

### Optimize for the temperature worst case

In this scenario the manufacturer would provide a recommended bias voltage for a temperature band, minimizing the power in that range. This approach is interesting because the user typically has more knowledge than the manufacturer about the expected environment for a given product. A sensor used in an office environment is unlikely to see temperatures of 85 °C and hence should be optimized around room temperature while the same device sold for industrial process monitoring may be exposed continuously to higher temperatures. The selection would be based on the worst case for each temperature point, selecting the static bias voltage for which the leakage power across all process corners is minimal. For the 85 °C case we fall back to the worst case where $V_{BB}$=0.6 V yields the lowest leakage power. However, for both −40 °C and 25 °C we find a better point as shown in Tab. 7.4

Table 7.2 – Optimizing for the Worst Case across all corners.

| | Case | SS | FS | TT | SF | FF |
|---|---|---|---|---|---|---|
| −40 °C | WC optimized | 14.5 nW | 11.7 nW | 17.2 nW | 41.4 nW | 27.7 nW |
| | adaptive $V_{BB}$ | 8.4 nW | 8.1 nW | 10.1 nW | 17.8 nW | 18.8 nW |
| | | 172.6% | 144.4% | 170.3% | 232.6% | 147.3% |
| 25 °C | WC optimized | 35.2 nW | 27.8 nW | 105.4 nW | 392.5 nW | 359.0 nW |
| | adaptive $V_{BB}$ | 24.9 nW | 27.7 nW | 71.4 nW | 292.5 nW | 283.0 nW |
| | | 141.4% | 100.4% | 147.6% | 134.2% | 126.8% |
| 85 °C | WC optimized | 198.2 nW | 243.3 nW | 539.6 nW | 1.7338 µW | 2.7081 µW |
| | adaptive $V_{BB}$ | 175.2 nW | 243.3 nW | 539.6 nW | 1.7338 µW | 2.7081 µW |
| | | 113.1% | 100% | 100% | 100% | 100% |

Table 7.3 – Optimisation for each corner lot worst case: Only gains for the SS corner.

| Case | SS −40 °C | SS 25 °C | SS 85 °C |
|---|---|---|---|
| SS opt. ($V_{BB}$ = −0.4 V) | 12.2 nW | 27.0 nW | 175.2 nW |
| rel. to WC opt. | 84.1% | 76.7% | 88.4% |
| rel. to adaptive | 145.2% | 108.4% | 100% |

Table 7.4 – Optimize separately for temperature: For 25 °C and −40 °C we do find a better retention point.

| | Case | SS | FS | TT | SF | FF |
|---|---|---|---|---|---|---|
| −40 °C | Temp. opt. ($V_{BB}$ = −0.1 V) | 9.1 nW | 8.1 nW | 10.9 nW | **20.3 nW** | 19.2 nW |
| | WC opt. ($V_{BB}$ = −0.6 V) | 14.5 nW | 11.7 nW | 17.2 nW | **41.4 nW** | 0.0277 nW |
| | rel. to WC opt. | 62.8% | 69.2% | 63.95% | 49.1% | 70.0% |
| | rel. to adaptive | 108.3% | 100% | 107.9% | 114.0% | 102.1% |
| 25 °C | Temp. opt. ($V_{BB}$ = −0.2 V) | 24.9 nW | 119.9 nW | 76.8 nW | 334.0 nW | **350.5 nW** |
| | WC opt. ($V_{BB}$ = −0.6 V) | 35.2 nW | 27.8 nW | 105.4 nW | **392.5 nW** | 359.0 nW |
| | relative to WC opt. | 70.7% | 431.3% | 72.9% | 85.1% | 97.6% |
| | relative to adaptive | 100% | 432.9% | 107.6% | 114.2% | 123.9% |

We observe in general an improvement with the worst case power for −40 °C reduced to 49.1% from 41.4 nW to 20.3 nW. Similarly, for the 25 °C case we realise a reduction of the worst case leakage power from 392.5 nW to 350.5 nW, corresponding to a reduction to 89.3%. However, the latter reduction comes at cost for the FS case where we observe a more than four fold increase in consumption which however still is far below the consumption seen in the FF corner.

## 7.6 Conclusion

In this chapter we have presented an alternative biasing approach, combining the concept of balanced currents with a charge pump directly operating on the wells of the bias domain. We presented a standard cell compatible current balance sensor, small enough to be distributed over the whole bias domain with a negligible overhead. Monte Carlo simulations show that the tripping point of the sensor stays within $\pm$ 100 mV with a Gaussian probability distribution, suggesting that variation of the sensor can be averaged out when integrating enough sensors. We show in simulation that the system in functional and capable to regulate the bias within an acceptable range. Finally, we show system power simulation numbers, including core leakage, charge pump power consumption, regulation overhead, sensor, and well leakage, realizing a ten-fold reduction of the leakage in the worst case corner while achieving a performance within 2.32 times of an ideal fully adaptive implementation across PVT.

# 8 Conclusion and Outlook

Typical applications for low power sensor nodes require battery powered autonomy over long periods of time while also providing a sufficiently high performance to evaluate state of the art data processing algorithms for local filtering. In order to combine these two contradicting requirements the system must be able to adapt, implementing a low power idle mode for long term retention as well as a fast operating mode for the occasional execution of demanding tasks.

Reducing the supply voltage is a well known technique to reduce circuit power, however comes with an increased susceptibility to process, voltage, and temperature (PVT) variation. We propose to use adaptive body biasing to compensate for the challenge imposed by variation. Moreover, we will use body biasing to optimize the same design for both high performance and idle mode despite the conflicting requirements.

For body biasing techniques to be effective, technology choice is key. The USJC 55 nm DDC technology which has been used throughout this thesis has a particularly strong body factor which opens up a design space where body biasing can compete with traditional low power methodologies such as adaptive voltage and frequency scaling.

First, we presented a methodology to exploit standard cell library characterisation data to compare cells timing variation across PVT corners, when applying body bias for PVT compensation. This methodology allows to judge the effectiveness of the body bias compensation technique. This insight is crucial for the designer to quantify the remaining uncertainty, needed to determine the margins required for a functional operation without timing violations. Comparing synthetic circuit realizations using a USJC 55n low power standard cell library, we showed that ABB is capable of compensating the PVT variation. ABB results in a reduction of the cross corner median delay variation by two orders of magnitude while also achieving a reduction of per-corner cell to cell variation.

Further we explored how standard cell characterisation data can be used as a tool for mapping the design space of adaptive supply and body voltage scaling. Designers can use this tool to define supply and bias conditions that will better suit their designs. A simple reference

design, implemented with a heavily pruned library, is used to sample the design space to derive leakage and dynamic energy maps. The analysis of the results shows how the minimum energy point can shift over a wide range when considering different operating frequencies and taking temperature and process variation into account, suggesting the need of a joint regulation of of body and supply. Further, we show that, using a single point calibration, these design space maps can be scaled to a more complex design utilizing the full library, yielding a model accuracy in the order of 1% of the energy report from a traditional signoff tool. Test chip measurements, employing bias and supply voltage predicted from the model for a constant frequency trajectory, result in a near perfect prediction of the dynamic energy within 5% and an acceptable match of the circuit leakage within a factor of 0.4 to 2 across temperature. Measurements of ring oscillators, constructed from a representative set of standard cells, show a frequency stability within 5% across the predicted constant frequency trajectory, validating the energy maps as an accurate modelling tool.

Next we presented an adaptive bias approach to overcome the impact of PVT variations on circuit frequency. The approach is based on the fact that circuit frequency is proportional to transistors on-currents. Regulating the on-currents relative to a programmable current reference by using body bias with the goal of keeping them constant in order to achieve a constant circuit frequency over PVT. Measurements of standard cell ring oscillators show a close match between on-current and frequency across process and temperature for forward biased conditions, with a tendency to slightly overcompensate for voltage. For a far reverse bias a worst case speed degradation of 40% and 60% is obtained across process and temperature, respectively. Core measurements yielded a similar degradation, suggesting that the worst cells of the critical path dominate the delay in far reverse, deep sub-threshold operation. The adaptive bias approach also allows to define different modes of operation of the circuit by defining the reference current that will allow the designer to control the performance-power trade-offs. Increasing the current will lead to faster designs with higher consumption, and decreasing the current will lead to slower designs with reduced consumption. Core leakage measurements show a 22-fold reduction when shifting the bias from a 10 MHz fast mode to a 310 kHz slow mode while the SRAM realizes a 59 times leakage reduction.

The second test chip adds an FLL on top of the on-current bias approach. Operation modes can now be set directly by defining a more meaningful target frequency that will define the on-current reference automatically. This approach also helps overcome the voltage overcompensation observed when fixing constant on-current reference that resulted in faster designs for low voltage and slower designs for high voltage conditions. The FLL includes a standard cell based programmable ring oscillator that is utilised to generate the system clock. The programmable oscillator length can be used to remove or add margin as necessary match the critical path length of the circuit to be compensated, and also allows to take the increased variation during the reverse slow operation into account. Core power measurements show an effective reduction of 320 times when comparing a fast operating mode at 8MHz and the idle retention mode.

Finally, we proposed a bias system focusing on slow modes of operation for low power moderate

performance applications. In this proposal the negative NMOS bias is regulated by a charge pump such that the on-current matches the PMOS which is directly supplied with a selectable positive bias. Heart of the system are multiple distributed, standard cell compatible, on-current balance sensors. Monte Carlo simulations yield a well controlled sensor switch behaviour with sufficient accuracy for a distributed system. The worst case oriented biasing approach shows that this simple bias system can be effective, achieving a ten fold leakage reduction for the worst case while realising a performance no worse than 2.32 times of a fully adaptive approach across the remaining PVT corners.

## Outlook

The strong body factor of 55 nm DDC left us with plenty of design space for exploiting supply voltage against body bias, however this would not be the case for other technologies where the less effective body factor might not even be sufficient to compensate process and temperature variations with body biasing. Hence, while the ideas are generally transferable, the constant on-current based biasing approach of Calanda and Nakayama is limited to technologies where the same on-currents can be achieved across the PVT range targeted for a particular device. However, the on-current based approach could in principle be extended to control an adaptive supply voltage and frequency scaling system if the target current is corrected for the well defined residual $1/V$ effect of the current on the frequency causing the overcompensation tendency within the first chip. Further research is needed to understand the tradeoffs, in particular if this kind of system is combined with adaptive body control.

Due to a layout mistake which has not been spotted with ERC, the biasing system in the final chip was not functional, resulting only in simulation results within this thesis. Hence, it would be nice to be able to fix the shorted wells and actually measure the current balanced control approach and validate the functionality in hardware. The approach is promising as it could be easily ported to other technologies as long as the NMOS and PMOS currents can be balanced with the biasing across PVT. Further, the approach could be combined with the secondary FLL loop by replacing the PMOS bias generator with a DAC, providing the user with a frequency knob.

The model based on the leakage- and dynamic energy maps could be integrated as part of a SoC modelling tool. It would be well suited for a discrete time simulator, replicating the different bias domains of a circuit as well as the interaction between them in different usage scenarios. This kind of simulator would allow to rapidly explore the design space for different SoC-Concepts and scenarios. When designing a new SoC, the model could be implemented first in a very basic form, estimating bias domain sizes and cell count as well as an activity. During the implementation process the model could be gradually updated, replacing the initial estimates with actual numbers from synthesis, place and route as well as from simulations. The scenarios could act as test cases similarly to the continuous integration approach in software engineering, but for power providing an early feedback if a specific design decision causes a

significant increase in overall system power consumption while also providing an insight where development effort should be steered to.

# Bibliography


[1] K. Ashton. (). That 'Internet of Things' Thing | RFID JOURNAL, [Online]. Available: https://www.rfidjournal.com/that-internet-of-things-thing (visited on 08/24/2020).

[2] V. Shnayder, M. Hempstead, B.-r. Chen, G. W. Allen, and M. Welsh, "Simulating the power consumption of large-scale sensor network applications", in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems*, ser. SenSys '04, Baltimore, MD, USA: Association for Computing Machinery, Nov. 3, 2004, pp. 188–200, ISBN: 978-1-58113-879-5. DOI: 10.1145/1031495.1031518. [Online]. Available: https://doi.org/10.1145/1031495.1031518 (visited on 09/04/2020).

[3] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI", in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2017, pp. 246–247. DOI: 10.1109/ISSCC.2017.7870353.

[4] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gürkaynak, and L. Benini, "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, pp. 2700–2713, Oct. 2017, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2017.2654506.

[5] S. G. Narendra and A. Chandrakasan, *Leakage in Nanometer CMOS Technologies (Series on Integrated Circuits and Systems)*. Berlin, Heidelberg: Springer-Verlag, 2005, ISBN: 978-0-387-25737-2.

[6] G. Tellez, A. Farrahi, and M. Sarrafzadeh, "Activity-driven clock design for low power circuits", in *Proceedings of IEEE International Conference on Computer Aided Design (ICCAD)*, Nov. 1995, pp. 62–65. DOI: 10.1109/ICCAD.1995.479992.

[7] Q. Wu, M. Pedram, and X. Wu, "Clock-gating and its application to low power design of sequential circuits", in *Proceedings of CICC 97 - Custom Integrated Circuits Conference*, May 1997, pp. 479–482. DOI: 10.1109/CICC.1997.606671.

[8] M. Samy Hosny and Y. Wu, "Low power clocking strategies in deep submicron technologies", in *2008 IEEE International Conference on Integrated Circuit Design and Technology and Tutorial*, Jun. 2008, pp. 143–146. DOI: 10.1109/ICICDT.2008.4567265.

[9] Jan M. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits: A Design Perspective*, 2nd ed., international ed., ser. Prentice Hall Electronics and VLSI Series. Upper Saddle River: Pearson Education International, 2003, 761 pp., ISBN: 978-0-13-120764-6.

[10] J. Rabaey, *Low Power Design Essentials*, ser. Integrated Circuits and Systems. Springer US, 2009, ISBN: 978-0-387-71712-8. DOI: 10.1007/978-0-387-71713-5. [Online]. Available: https://www.springer.com/de/book/9780387717128 (visited on 08/07/2020).

[11] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 299–316, Jun. 2000, ISSN: 1557-9999. DOI: 10.1109/92.845896.

[12] D. Bol, J. de Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd, D. Flandre, and J. Legat, "SleepWalker: A 25-MHz 0.4-V Sub- 7- Microcontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes", *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 1, pp. 20–32, Jan. 1, 2013. DOI: 10.1109/JSSC.2012.2218067. [Online]. Available: http://ieeexplore.ieee.org/document/6332542/.

[13] J. Myers, A. Savanth, D. Howard, R. Gaddh, P. Prabhat, and D. Flynn, "8.1 An 80nW retention 11.7pJ/cycle active subthreshold ARM Cortex-M0+ subsystem in 65nm CMOS for WSN applications", in *2015 IEEE International Solid- State Circuits Conference - (ISSCC)*, IEEE, Jan. 1, 2015, pp. 1–3, ISBN: 978-1-4799-6223-5. DOI: 10.1109/ISSCC.2015.7062967. [Online]. Available: http://ieeexplore.ieee.org/document/7062967/.

[14] B. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits", in *Proceedings of the 2004 International Symposium on Low Power Electronics and Design (IEEE Cat. No.04TH8758)*, Aug. 2004, pp. 90–95. DOI: 10.1109/LPE.2004.240808.

[15] R. G. Dreslinski, M. Wieckowski, D. T. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits", presented at the Proceedings of the IEEE, vol. 98, Jan. 1, 2010, pp. 253–266. DOI: 10.1109/JPROC.2009.2034764. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5395763.

[16] B. Keller, M. Cochet, B. Zimmer, Y. Lee, M. Blagojevic, J. Kwak, A. Puggelli, S. Bailey, P.-F. Chiu, P. Dabbelt, C. Schmidt, E. Alon, K. Asanović, and B. Nikolic, "Sub-microsecond adaptive voltage scaling in a 28nm FD-SOI processor SoC", in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, IEEE, Jan. 1, 2016, pp. 269–272, ISBN: 978-1-5090-2972-3. DOI: 10.1109/ESSCIRC.2016.7598294. [Online]. Available: http://ieeexplore.ieee.org/document/7598294/.

[17] J. Constantin, A. Bonetti, A. Teman, C. Müller, L. Schmid, and A. Burg, "DynOR: A 32-bit microprocessor in 28 nm FD-SOI with cycle-by-cycle dynamic clock adjustment", in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Sep. 2016, pp. 261–264. DOI: 10.1109/ESSCIRC.2016.7598292.

[18] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage", *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 1, 2002. DOI: 10.1109/JSSC.2002.803949. [Online]. Available: http://ieeexplore.ieee.org/document/1046081/.

[19] M. Bhushan, A. Gattiker, M. Ketchen, and K. Das, "Ring oscillators for CMOS process tuning and variability control", *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10–18, Feb. 2006, ISSN: 1558-2345. DOI: 10.1109/TSM.2005.863244.

[20] Q. Liu and S. S. Sapatnekar, "Capturing Post-Silicon Variations Using a Representative Critical Path", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 2, pp. 211–222, Feb. 2010, ISSN: 1937-4151. DOI: 10.1109/TCAD.2009.2035552.

[21] J. Park and J. A. Abraham, "A fast, accurate and simple critical path monitor for improving energy-delay product in DVS systems", in *IEEE/ACM International Symposium on Low Power Electronics and Design*, Aug. 2011, pp. 391–396. DOI: 10.1109/ISLPED.2011.5993672.

[22] D. Ernst, Nam Sung Kim, S. Das, S. Pant, R. Rao, Toan Pham, C. Ziesler, D. T. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: A low-power pipeline based on circuit-level timing speculation", in *MICRO-03*, IEEE Comput. Soc, Jan. 1, 2003, pp. 7–18, ISBN: 0-7695-2043-X. DOI: 10.1109/MICRO.2003.1253179. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1253179.

[23] S. Das, C. Tokunaga, S. Pant, W.-H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, "RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance", *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 1, 2009. DOI: 10.1109/JSSC.2008.2007145. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4735568.

[24] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. M. Harris, D. T. Blaauw, and D. Sylvester, "Bubble Razor: Eliminating Timing Margins in an ARM Cortex-M3 Processor in 45 nm CMOS Using Architecturally Independent Error Detection and Correction", *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 66–81, Jan. 1, 2013. DOI: 10.1109/JSSC.2012.2220912. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6365269.

[25] I. Kwon, S. Kim, D. Fick, M. Kim, Y.-P. Chen, and D. Sylvester, "Razor-Lite: A Light-Weight Register for Error Detection by Observing Virtual Supply Rails", *IEEE Journal of Solid-State Circuits*, vol. 49, no. 9, pp. 2054–2066, Jan. 1, 2014. DOI: 10.1109/JSSC.2014.2328658. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6842679.

[26] Y. Zhang, M. Khayatzadeh, K. Yang, M. Saligane, N. Pinckney, M. Alioto, D. T. Blaauw, and D. Sylvester, "8.8 iRazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor", in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, Jan. 1, 2016, pp. 160–162, ISBN: 978-1-4673-9466-6. DOI: 10.1109/ISSCC.2016.7417956. [Online]. Available: http://ieeexplore.ieee.org/document/7417956/.

[27] C. Gallon, C. Fenouillet-Beranger, A. Vandooren, F. Boeuf, S. Monfray, F. Payet, S. Orain, V. Fiori, F. Salvetti, N. Loubet, C. Charbuillet, A. Toffoli, F. Allain, K. Romanjek, I. Cayrefourcq, B. Ghyselen, C. Mazure, D. Delille, F. Judong, C. Perrot, M. Hopstaken, P. Scheblin, P. Rivallin, L. Brevard, O. Faynot, S. Cristoloveanu, and T. Skotnicki, "Ultra-Thin Fully Depleted SOI Devices with Thin BOX, Ground Plane and Strained Liner Booster", in *2006 IEEE International SOI Conferencee Proceedings*, Oct. 2006, pp. 17–18. DOI: 10.1109/SOI.2006.284410.

[28] L. T. Clark, D. Zhao, T. Bakhishev, H. Ahn, E. Boling, M. Duane, K. Fujita, P. Gregory, T. Hoffmann, M. Hori, D. Kanai, D. Kidd, S. Lee, Y. Liu, J. Mitani, J. Nagayama, S. Pradhan, P. Ranade, R. Rogenmoser, L. Scudder, L. Shifren, Y. Torii, M. Wojko, Y. Asada, T. Ema, and S. Thompson, "A highly integrated 65-nm SoC process with enhanced power/performance of digital and analog circuits", in *2012 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Jan. 1, 2012, pp. 14.4.1–14.4.4, ISBN: 0163-1918. DOI: 10.1109/IEDM.2012.6479042. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6479042.

[29] H. N. Patel, A. Roy, F. B. Yahya, N. Liu, B. Calhoun, K. Kumeno, M. Yasuda, A. Harada, and T. Ema, "A 55nm Ultra Low Leakage Deeply Depleted Channel technology optimized for energy minimization in subthreshold SRAM and logic", in *ESSDERC 2016 - 46th European Solid-State Device Research Conference*, IEEE, Jan. 1, 2016, pp. 37–40, ISBN: 978-1-5090-2969-3. DOI: 10.1109/ESSDERC.2016.7599583. [Online]. Available: http://ieeexplore.ieee.org/document/7599583/.

[30] K. Fujita, Y. Torii, M. Hori, J. Oh, L. Shifren, P. Ranade, M. Nakagawa, K. Okabe, T. Miyake, K. Ohkoshi, M. Kuramae, T. Mori, T. Tsuruta, S. Thompson, and T. Ema, "Advanced channel engineering achieving aggressive reduction of VT variation for ultra-low-power applications", in *2011 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Jan. 1, 2011, pp. 32.3.1–32.3.4, ISBN: 978-1-4577-0506-9. DOI: 10.1109/IEDM.2011.6131657. [Online]. Available: http://ieeexplore.ieee.org/document/6131657/.

[31] F. Andrieu, O. Weber, S. Baudot, C. Fenouillet-Béranger, O. Rozeau, J. Mazurier, P. Perreau, J. Eymery, and O. Faynot, "Fully depleted Silicon-On-Insulator with back bias and strain for low power and high performance applications", in *2010 IEEE International Conference on Integrated Circuit Design and Technology*, Jun. 2010, pp. 59–62. DOI: 10.1109/ICICDT.2010.5510295.

[32] T. Skotnicki and S. Monfray, "UTBB FDSOI: Evolution and opportunities", in *2015 45th European Solid State Device Research Conference (ESSDERC)*, Sep. 2015, pp. 76–79. DOI: 10.1109/ESSDERC.2015.7324717.

[33] E. Beigne, A. Valentian, B. Giraud, O. Thomas, T. Benoist, Y. Thonnart, S. Bernard, G. Moritz, O. Billoint, Y. Maneglia, P. Flatresse, J. P. Noel, F. Abouzeid, B. Pelloux-Prayer, A. Grover, S. Clerc, P. Roche, J. Le Coz, S. Engels, and R. Wilson, "Ultra-Wide Voltage Range designs in Fully-Depleted Silicon-On-Insulator FETs", presented at the Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013, IEEE Conference

Publications, Jan. 1, 2013, pp. 613–618, ISBN: 978-1-4673-5071-6. DOI: 10.7873/DATE.2013.135. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6513581.

[34] A. Quelen, G. Pillonnet, P. Flatresse, and E. Beigné, "A 2.5uW 0.0067mm2 automatic back-biasing compensation unit achieving 50% leakage reduction in FDSOI 28nm over 0.35-to-1V VDDrange", in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb. 2018, pp. 304–306. DOI: 10.1109/ISSCC.2018.8310305.

[35] M. Pons, C. T. Müller, D. Ruffieux, J.-L. Nagel, S. Emery, A. Burg, S. Tanahashi, Y. Tanaka, and A. Takeuchi, "A 0.5 V 2.5 $\mu$W/MHz Microcontroller with Analog-Assisted Adaptive Body Bias PVT Compensation with 3.13nW/kB SRAM Retention in 55nm Deeply-Depleted Channel CMOS", in *2019 IEEE Custom Integrated Circuits Conference (CICC)*, Apr. 2019, pp. 1–4. DOI: 10.1109/CICC.2019.8780199.

[36] T. C. Müller, J.-L. Nagel, M. Pons, D. Séverac, K. Hashiba, S. Sawada, K. Miyatake, S. Emery, and A. Burg, "PVT compensation in Mie Fujitsu 55 nm DDC: A standard-cell library based comparison", in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2017, pp. 1–2. DOI: 10.1109/S3S.2017.8309246.

[37] C. T. Müller, M. Pons, D. Ruffieux, J.-L. Nagel, S. Emery, A. Burg, S. Tanahashi, Y. Tanaka, and A. Takeuchi, "Minimum Energy Point in Constant Frequency Designs under Adaptive Supply Voltage and Body Bias Adjustment in 55 nm DDC", in *2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, Jul. 2019, pp. 285–288. DOI: 10.1109/PRIME.2019.8787736.

[38] S. Jain, L. Lin, and M. A. T. o. C. a. S. I, "Design-Oriented Energy Models for Wide Voltage Scaling Down to the Minimum Energy Point", *Procedings of ISSCC 2015*, 2017. DOI: 10.1109/TCSI.2017.2736540. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/8013083/.

[39] O. Andersson, S. M. Y. Sherazi, and J. N. Rodrigues, "Impact of switching activity on the energy minimum voltage for 65 nm sub-VT CMOS", *NORCHIP, 2011*, pp. 1–4, Jan. 1, 2011, ISSN: 978-1-4577-0514-4. DOI: 10.1109/NORCHP.2011.6126748. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6126748.

[40] M. Miyazaki, J. Kao, and A. P. Chandrakasan, "A 175mV multiply-accumulate unit using an adaptive supply voltage and body bias (ASB) architecture", in *2002 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (Cat. No.02CH37315)*, vol. 2, IEEE, Jan. 1, 2002, pp. 40–391, ISBN: 0-7803-7335-9. DOI: 10.1109/ISSCC.2002.992099. [Online]. Available: http://ieeexplore.ieee.org/document/992099/.

[41] J. Lee, Y. Zhang, Q. Dong, W. Lim, M. Saligane, Y. Kim, S. Jeong, J. Lim, M. Yasuda, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, "19.2 A 6.4pJ/Cycle Self-Tuning Cortex-M0 IoT Processor Based on Leakage-Ratio Measurement for Energy-Optimal Operation Across Wide-Range PVT Variation", in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, Feb. 2019, pp. 314–315. DOI: 10.1109/ISSCC.2019.8662454.

[42] S. S. Sapatnekar, "What happens when circuits grow old: Aging issues in CMOS design", in *2013 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, Apr. 2013, pp. 1–2. DOI: 10.1109/VLSI-TSA.2013.6545621.

[43] J. Daga, E. Ottaviano, and D. Auvergne, "Temperature effect on delay for low voltage applications [CMOS ICs]", in *Proceedings Design, Automation and Test in Europe*, Feb. 1998, pp. 680–685. DOI: 10.1109/DATE.1998.655931.

[44] S. Borkar and A. A. Chien, "The future of microprocessors", *Commun. ACM*, vol. 54, no. 5, pp. 67–77, May 1, 2011, ISSN: 0001-0782. DOI: 10.1145/1941487.1941507. [Online]. Available: https://doi.org/10.1145/1941487.1941507 (visited on 10/12/2020).

[45] M. B. Taylor, "A Landscape of the New Dark Silicon Design Regime", *IEEE Micro*, vol. 33, no. 5, pp. 8–19, Sep. 2013, ISSN: 1937-4143. DOI: 10.1109/MM.2013.90.

[46] J. Lee, Y. Zhang, Q. Dong, W. Lim, M. Saligane, Y. Kim, S. Jeong, J. Lim, M. Yasuda, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, "A Self-Tuning IoT Processor Using Leakage-Ratio Measurement for Energy-Optimal Operation", *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 87–97, Jan. 2020, ISSN: 1558-173X. DOI: 10.1109/JSSC.2019.2939890.

[47] M. Meijer, J. P. de Gyvez, B. Kup, B. van Uden, P. Bastiaansen, M. Lammers, and M. Vertregt, "A forward body bias generator for digital CMOS circuits with supply voltage scaling", in *2010 IEEE International Symposium on Circuits and Systems - ISCAS 2010*, IEEE, Jan. 1, 2010, pp. 2482–2485, ISBN: 978-1-4244-5308-5. DOI: 10.1109/ISCAS.2010.5537129. [Online]. Available: http://ieeexplore.ieee.org/document/5537129/.

[48] M. Blagojević, M. Cochet, B. Keller, P. Flatresse, A. Vladimirescu, and B. Nikolić, "A fast, flexible, positive and negative adaptive body-bias generator in 28nm FDSOI", in *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, Jun. 2016, pp. 1–2. DOI: 10.1109/VLSIC.2016.7573479.

[49] F. Abouzeid, C. Bernicot, S. Clerc, J.-M. Daveau, G. Gasiot, D. Noblet, D. Soussan, and P. Roche, "30% Static Power Improvement on ARM Cortex-A53 using Static Biasing-Anticipation", presented at the 2016 ESSCIRC, Apr. 25, 2016, pp. 1–4.

[50] D. Rossi, I. Loi, A. Pullini, C. Müller, A. Burg, F. Conti, L. Benini, and P. Flatresse, "A Self-Aware Architecture for PVT Compensation and Power Nap in Near Threshold Processors", *IEEE Design & Test*, vol. 34, no. 6, pp. 46–53, Jan. 1, 2017. DOI: 10.1109/MDAT.2017.2750907. [Online]. Available: http://ieeexplore.ieee.org/document/8031073/.

[51] S. Martin, K. Flautner, T. Mudge, and D. Blaauw, "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads", in *IEEE/ACM International Conference on Computer Aided Design, 2002. ICCAD 2002.*, Nov. 2002, pp. 721–725. DOI: 10.1109/ICCAD.2002.1167611.

[52] S. H. Kulkarni, D. M. Sylvester, and D. T. Blaauw, "Design-Time Optimization of Post-Silicon Tuned Circuits Using Adaptive Body Bias", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 3, pp. 481–494, 2008. DOI: 10.1109/TCAD.2008.915529. [Online]. Available: http://ieeexplore.ieee.org/document/4454014/.

[53] J. M. Kühn, H. Amano, O. Bringmann, and W. Rosenstiel, "Leveraging FDSOI through body bias domain partitioning and bias search", in *DAC '16*, ACM Press, Jan. 1, 2016, pp. 1–6, ISBN: 978-1-4503-4236-0. DOI: 10.1145/2897937.2898039. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2897937.2898039.

[54] L. Lin, S. Jain, and M. A. S.-S. C. Conference, "A 595pW 14pJ/Cycle microcontroller with dual-mode standard cells and self-startup for battery-indifferent distributed sensing", *Procedings of ISSCC 2018*, 2018. DOI: 10.1109/ISSCC.2018.8310175","publicationTitle":"Solid. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8310175/.

[55] H. Reyserhove and W. Dehaene, "A Differential Transmission Gate Design Flow for Minimum Energy Sub-10-pJ/Cycle ARM Cortex-M0 MCUs", *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1904–1914, Jul. 2017, ISSN: 0018-9200. DOI: 10.1109/JSSC.2017.2693241.

[56] M. Hiienkari, N. Gupta, J. Teittinen, J. Simonsson, M. Turnquist, J. Eriksson, R. Anttila, O. Myllynen, H. Rämäkkö, S. Mäkikyrö, and L. Koskinen, "A 0.4-0.9V, 2.87pJ/cycle Near-Threshold ARM Cortex-M3 CPU with In-Situ Monitoring and Adaptive-Logic Scan", in *2020 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, Apr. 2020, pp. 1–3. DOI: 10.1109/COOLCHIPS49199.2020.9097634.

[57] J.-L. Nagel, S. M. Pons, A.-S. Porret, D. Ruffieux, G. C. A. Salazar, and D. Séverac, "Compensation Device for Compensating Pvt Variations of an Analog and/or Digital Circuit", European pat. 3488527A1, May 29, 2019.

[58] C. Arm, S. Gyger, J.-M. Masgonty, M. Morgan, J.-L. Nagel, C. Piguet, F. Rampogna, and P. Volet, "Low-power 32-bit dual-MAC 120 $\mu$W/MHz 1.0 V icyflex DSP/MCU core", in *ESSCIRC 2008 - 34th European Solid-State Circuits Conference*, Sep. 2008, pp. 190–193. DOI: 10.1109/ESSCIRC.2008.4681824.

[59] L. T. Clark, D. Kidd, V. Agrawal, S. Leshner, and G. Krishnan, "Independent N and P process monitors for body bias based process corner correction", in *2014 IEEE Custom Integrated Circuits Conference - CICC 2014*, IEEE, Jan. 1, 2014, pp. 1–4, ISBN: 978-1-4799-3286-3. DOI: 10.1109/CICC.2014.6946092. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6946092.

[60] A. A. Vatanjou, T. Ytterdal, and S. Aunet, "28 nm UTBB-FDSOI energy efficient and variation tolerant custom digital-cell library with application to a subthreshold MAC block", in *2016 MIXDES - 23rd International Conference Mixed Design of Integrated Circuits and Systems*, Jun. 2016, pp. 105–110. DOI: 10.1109/MIXDES.2016.7529711.

[61] J. V. De la Cruz and A. L. Aita, "A 1-V PTAT current reference circuit with 0.05%/V current sensitivity to VDD", in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 502–505. DOI: 10.1109/ISCAS.2016.7527287.

[62] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, Jan. 20, 2009, 896 pp., ISBN: 978-0-470-24599-6.

[63] F. Maloberti, *Data Converters*. Dordrecht, Netherlands: Springer, 2007, 440 pp., ISBN: 978-0-387-32485-2.

[64] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Transistor Variability Modeling and its Validation With Ring-Oscillation Frequencies for Body-Biased Subthreshold Circuits", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 7, pp. 1118–1129, Jul. 2010, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2009.2020594.

[65] R. G. Gomez, E. Bano, and S. Clerc, "Comparative evaluation of Body Biasing and Voltage Scaling for Low-Power Design on 28nm UTBB FD-SOI Technology", in *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Jul. 2019, pp. 1–6. DOI: 10.1109/ISLPED.2019.8824791.

[66] M. Cochet, B. Pelloux-Prayer, M. Saligane, S. Clerc, P. Roche, J. Autran, and D. Sylvester, "Experimental model of adaptive body biasing for energy efficiency in 28nm UTBB FD-SOI", in *2014 SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2014, pp. 1–2. DOI: 10.1109/S3S.2014.7028221.

[67] PyVISA Authors, *PyVISA*. [Online]. Available: https://pyvisa.readthedocs.io/en/stable/.

[68] OpenOCD Authors, *OpenOCD*. [Online]. Available: http://openocd.org.

[69] D. Rath, "Design and Implementation of an On-Chip Debug Solution for Embedded Target Systems based on the ARM7 and ARM9 Family", Diploma Thesis, University of Applied Sciences Augsburg, Augsburg.

[70] N. Misawa, H. Kurata, K. Kumeno, R. Nanjo, M. Kai, T. Ema, and M. P. Sole, "SNM analytical approach to robust subthreshold SRAM operation based on the 55nm DDC technology", in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2017, pp. 1–2. DOI: 10.1109/S3S.2017.8308741.

[71] T. Fukuda, K. Kohara, T. Dozaka, Y. Takeyama, T. Midorikawa, K. Hashimoto, I. Wakiyama, S. Miyano, and T. Hojo, "13.4 A 7ns-access-time 25$\mu$W/MHz 128kb SRAM for low-power fast wake-up MCU in 65nm CMOS with 27fA/b retention current", in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb. 2014, pp. 236–237. DOI: 10.1109/ISSCC.2014.6757415.

[72] CSEM SA, *Icyflex-V*. [Online]. Available: https://www.csem.ch/pdf/51525?name=icyflex-v-web.pdf.

[73] J.-L. Nagel, C. Arm, R. Cattenoz, H.-R. Graf, and V. Moser, "Icyflex-V—a New Ultra-low-power Processor based on RISC-V Architecture", *CSEM Scientific and Technical Report*, no. 2019, p. 116, [Online]. Available: https://www.csem.ch/pdf/128515?name=CSEM-STR-2019-p116.pdf.

[74] CSEM SA, *icyTRX Ultra low-power 2.4GHz transceiver for Bluetooth 5.2, 802.15.4 & IoT*. [Online]. Available: https://www.csem.ch/Doc.aspx?id=41379.

[75] Y. Zhang, N. Suda, L. Lai, and V. Chandra. (Feb. 14, 2018). Hello Edge: Keyword Spotting on Microcontrollers, [Online]. Available: http://arxiv.org/abs/1711.07128 (visited on 09/22/2020).

[76] L. Lai, N. Suda, and V. Chandra. (Jan. 19, 2018). CMSIS-NN: Efficient Neural Network Kernels for Arm Cortex-M CPUs, [Online]. Available: http://arxiv.org/abs/1801.06601 (visited on 09/22/2020).

[77] E. Torti, A. Fontanella, M. Musci, N. Blago, D. Pau, F. Leporati, and M. Piastra, "Embedded Real-Time Fall Detection with Deep Learning on Wearable Devices", in *2018 21st Euromicro Conference on Digital System Design (DSD)*, Aug. 2018, pp. 405–412. DOI: 10.1109/DSD.2018.00075.

[78] C. Albea-Sanchez, D. Puschini, S. Lesecq, E. Beigné, and P. Vivet, "Architecture and Control of a Digital Frequency-Locked Loop for Fine-Grain Dynamic Voltage and Frequency Scaling in Globally Asynchronous Locally Synchronous Structures", *Journal of Low Power Electronics*, vol. 7, no. 3, 328–340(13), Aug. 2011. DOI: 10.1166/jolpe.2011.1141. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00255726 (visited on 09/23/2020).

[79] I. Miro-Panades, E. Beigné, Y. Thonnart, L. Alacoque, P. Vivet, S. Lesecq, D. Puschini, A. Molnos, F. Thabet, B. Tain, K. Ben Chehida, S. Engels, R. Wilson, and D. Fuin, "A Fine-Grain Variation-Aware Dynamic $\rm Vdd$-Hopping AVFS Architecture on a 32 nm GALS MPSoC", *IEEE Journal of Solid-State Circuits*, vol. 49, no. 7, pp. 1475–1486, Jul. 2014, ISSN: 1558-173X. DOI: 10.1109/JSSC.2014.2317137.

[80] B. Calhoun and A. Chandrakasan, "Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm CMOS", in *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, Feb. 2005, 300–599 Vol. 1. DOI: 10.1109/ISSCC.2005.1493988.

[81] T. V. Breussegem and M. Steyaert, *CMOS Integrated Capacitive DC-DC Converters*, ser. Analog Circuits and Signal Processing. New York: Springer-Verlag, 2013, ISBN: 978-1-4614-4279-0. DOI: 10.1007/978-1-4614-4280-6. [Online]. Available: https://www.springer.com/gp/book/9781461442790 (visited on 10/12/2020).

[82] C. Müller, S. Malkowsky, O. Andersson, B. Mohammadi, J. Sparsø, and J. N. Rodrigues, "A 65-nm CMOS area optimized de-synchronization flow for sub-VT designs", in *2013 IFIP/IEEE 21st International Conference on Very Large Scale Integration (VLSI-SoC)*, Oct. 2013, pp. 380–385. DOI: 10.1109/VLSI-SoC.2013.6673313.

[83] C. T. Müller, E. Kasapaki, R. B. Sørensen, and J. Sparsø, "Synthesis and layout of an asynchronous network-on-chip using Standard EDA tools", in *2014 NORCHIP*, Oct. 2014, pp. 1–6. DOI: 10.1109/NORCHIP.2014.7004742.

[84] E. Kasapaki, M. Schoeberl, R. B. Sørensen, C. Müller, K. Goossens, and J. Sparsø, "Argo: A Real-Time Network-on-Chip Architecture With an Efficient GALS Implementation", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 2, pp. 479–492, Feb. 2016, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2015.2405614.

[85] P. Giard, A. Balatsoukas-Stimming, T. C. Müller, A. Burg, C. Thibeault, and W. J. Gross, "A multi-Gbps unrolled hardware list decoder for a systematic polar code", in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 1194–1198. DOI: 10.1109/ACSSC.2016.7869561.

[86] P. Giard, A. Balatsoukas-Stimming, T. C. Müller, A. Bonetti, C. Thibeault, W. J. Gross, P. Flatresse, and A. Burg, "PolarBear: A 28-nm FD-SOI ASIC for Decoding of Polar Codes", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 616–629, Dec. 2017, ISSN: 2156-3365. DOI: 10.1109/JETCAS.2017.2745704.

[87] R. Ghanaatian, A. Balatsoukas-Stimming, T. C. Müller, M. Meidlinger, G. Matz, A. Teman, and A. Burg, "A 588-Gb/s LDPC Decoder Based on Finite-Alphabet Message Passing", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 2, pp. 329–340, Feb. 2018, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2017.2766925.

# List of Figures

# List of Tables

# Curriculum Vitae


| | |
|---|---|
| Name: | **Thomas Christoph MÜLLER** |
| Date of Birth: | 26.02.1986 |
| Nationality: | German |
| Address: | EPFL, STI-IEL-TCL |
| | Station 11 |
| | CH-1015 Lausanne |
| | Switzerland |
| E-Mail: | christoph.mueller@epfl.ch |
| | christoph@christophmueller.org |


## Education

| | |
|---|---|
| 04.2015 – present | **École Polytechnique Fédérale de Lausanne**, Lausanne (VD), CH<br>Ph.D Degree in Electrical Engineering |
| 10.2009 – 04.2012 | **Lund University**, Lund, SE<br>Master's Degree in System on Chip |
| 09.2006 – 06.2011 | **University of Applied Sciences Schmalkalden**, Schmalkalden, DE<br>Bachelor's Degree in Information Technology |

## Professional Experience

| | |
|---|---|
| 04.2015 – present | **École Polytechnique Fédérale de Lausanne**, Lausanne (VD), CH<br>Doctoral Assistant at Telecommunications Circuits Laboratory |
| 07.2016 – 09.2020 | **CSEM SA**, Neuchâtel (NE), CH<br>PhD Student in the System on Chip division |

| | |
|---|---|
| 09.2014 – 10.2014 | **Lund University**, Lund, SE<br>Project Assistant |
| 11.2013 – 08.2014 | **Technical University of Denmark (DTU)**, Kongens Lyngby, DK<br>Research Assistant |
| 08.2013 – 11.2014 | **Lund University**, Lund, SE<br>Project Assistant |
| 09.2010 – 06.2011 | **Fraunhofer Institute for Integrated Circuits IIS**, Erlangen, DE<br>Internship and Bachelor Thesis |

# List of Publications

## Part of this Thesis

T. C. Müller, J.-L. Nagel, M. Pons, D. Séverac, K. Hashiba, S. Sawada, K. Miyatake, S. Emery, and A. Burg, "PVT compensation in Mie Fujitsu 55 nm DDC: A standard-cell library based comparison", in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2017, pp. 1–2. DOI: 10.1109/S3S.2017.8309246

C. T. Müller, M. Pons, D. Ruffieux, J.-L. Nagel, S. Emery, A. Burg, S. Tanahashi, Y. Tanaka, and A. Takeuchi, "Minimum Energy Point in Constant Frequency Designs under Adaptive Supply Voltage and Body Bias Adjustment in 55 nm DDC", in *2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, Jul. 2019, pp. 285–288. DOI: 10.1109/PRIME.2019.8787736

M. Pons, C. T. Müller, D. Ruffieux, J.-L. Nagel, S. Emery, A. Burg, S. Tanahashi, Y. Tanaka, and A. Takeuchi, "A 0.5 V 2.5 $\mu$W/MHz Microcontroller with Analog-Assisted Adaptive Body Bias PVT Compensation with 3.13nW/kB SRAM Retention in 55nm Deeply-Depleted Channel CMOS", in *2019 IEEE Custom Integrated Circuits Conference (CICC)*, Apr. 2019, pp. 1–4. DOI: 10.1109/CICC.2019.8780199

## Outside the Scope of this Thesis

C. Müller, S. Malkowsky, O. Andersson, B. Mohammadi, J. Sparsø, and J. N. Rodrigues, "A 65-nm CMOS area optimized de-synchronization flow for sub-VT designs", in *2013 IFIP/IEEE 21st International Conference on Very Large Scale Integration (VLSI-SoC)*, Oct. 2013, pp. 380–385. DOI: 10.1109/VLSI-SoC.2013.6673313

C. T. Müller, E. Kasapaki, R. B. Sørensen, and J. Sparsø, "Synthesis and layout of an asynchronous network-on-chip using Standard EDA tools", in *2014 NORCHIP*, Oct. 2014, pp. 1–6. DOI: 10.1109/NORCHIP.2014.7004742

E. Kasapaki, M. Schoeberl, R. B. Sørensen, C. Müller, K. Goossens, and J. Sparsø, "Argo: A Real-Time Network-on-Chip Architecture With an Efficient GALS Implementation", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 2, pp. 479–492, Feb. 2016, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2015.2405614

J. Constantin, A. Bonetti, A. Teman, C. Müller, L. Schmid, and A. Burg, "DynOR: A 32-bit microprocessor in 28 nm FD-SOI with cycle-by-cycle dynamic clock adjustment", in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Sep. 2016, pp. 261–264. DOI: 10.1109/ESSCIRC.2016.7598292

P. Giard, A. Balatsoukas-Stimming, T. C. Müller, A. Burg, C. Thibeault, and W. J. Gross, "A multi-Gbps unrolled hardware list decoder for a systematic polar code", in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 1194–1198. DOI: 10.1109/ACSSC.2016.7869561

D. Rossi, I. Loi, A. Pullini, C. Müller, A. Burg, F. Conti, L. Benini, and P. Flatresse, "A Self-Aware Architecture for PVT Compensation and Power Nap in Near Threshold Processors", *IEEE Design & Test*, vol. 34, no. 6, pp. 46–53, Jan. 1, 2017. DOI: 10.1109/MDAT.2017.2750907. [Online]. Available: http://ieeexplore.ieee.org/document/8031073/

P. Giard, A. Balatsoukas-Stimming, T. C. Müller, A. Bonetti, C. Thibeault, W. J. Gross, P. Flatresse, and A. Burg, "PolarBear: A 28-nm FD-SOI ASIC for Decoding of Polar Codes", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 616–629, Dec. 2017, ISSN: 2156-3365. DOI: 10.1109/JETCAS.2017.2745704

R. Ghanaatian, A. Balatsoukas-Stimming, T. C. Müller, M. Meidlinger, G. Matz, A. Teman, and A. Burg, "A 588-Gb/s LDPC Decoder Based on Finite-Alphabet Message Passing", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 2, pp. 329–340, Feb. 2018, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2017.2766925