MONITOR AND CONTROL STRATEGIES TO REDUCE THE IMPACT OF PROCESS VARIATIONS IN DIGITAL CIRCUITS

Joan Mauricio Ferré
PhD Candidate

Francesc Moll Echeto
Thesis Director

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Enginyeria Electrònica

Thesis submitted in partial fulfillment of the requirement for the

PhD Degree issued by the Universitat Politècnica de Catalunya,

in its Electronic Engineering Program.

September 2015

Quan hi ha temporal, què és més fort, el jonc o el roure? El roure s'alça imponent, i enorme i orgullós combat el vent amb la seva gruixuda escorça. El jonc manca de mida i escorça, la blana tija no pot lluitar de la mateixa manera. L'un resisteix, immutable. I de vegades venç, quan el vent es rendeix. Però d'altres, la seva rigidesa el converteix en un amàs d'estelles. El roure guanya de vegades. L'altre mai no resisteix i sempre ho fa. Quan el vent bufa i l'amenaça, no el combat, sinó que s'inclina en la direcció que aquest marca i usa la seva força al seu favor. La seva flexibilitat fa que es doblegui, sense arribar mai a trencar-se. El jonc guanya sempre.

Proverbi xinès.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

As CMOS technology scales down, Process, Voltage, Temperature and Ageing (PVTA) variations have an increasing impact on the performance and power consumption of electronic devices. These issues may hold back the continuous improvement of these devices in the near future. There are several ways to face the variability problem: to increase the operating margins of maximum clock frequency, the implementation of lithographic friendly layout styles, and the last one and the focus of this thesis, to adapt the circuit to its actual manufacturing and environment conditions by tuning some of the adjustable parameters once the circuit has been manufactured. The main challenge of this thesis is to develop a low-area variability compensation mechanism to automatically mitigate PVTA variations in run-time, i.e. while integrated circuit is running. This implies the development of a sensor to obtain the most accurate picture of variability, and the implementation of a control block to knob some of the electrical parameters of the circuit.

ACKNOWLEDGMENTS

This work was partly supported by the European Community Seventh Framework Programme (FP7/2007-2013) under grant agreement number 248538 (MODERN project) and the Spanish Ministry of Economy (MINECO) and FEDER funds through project TEC2008-01856 (Terasystems) and TEC2013-45638-C3-2-R (Maragda).

ACRONYMS

ABB	Adaptive Body Bias
AVS	Adaptive Voltage Scaling
BBG	Body Bias Generator
BBI	Body Bias Island
BTI	Bias Temperature Instability
CDF	Cummulative Distribution Function
CIC	Cascaded Integrator Comb
CMOS	Complementary MOSFET
CPU	Central Process Unit
D2D	Die-to-Die
DAC	Digital-to-Analog Converter
DFM	Design For Manufacturability
EDA	Electronic Design Automation
EDS	Error-Detection Sequential
F-F	Fast-Fast
FBB	Forward Body Bias
FC	Full Custom
FD-SOI	Fully Depleted Silicon on Insulator
FF	Flip-Flop
FGBB	Fine Grain Body Bias

FinFET	Fin Field Effect Transistor
HCI	Hot-Carrier-Induced
HIPICS	High Performance Integrated Circuits and Systems Design Group
HPM	Hardware Performance Monitor
IC	Integrated Circuits
INL	Integral Non Linearity
IR	Interface Roughness
IR	Infrared
LAVS	Local Adaptive Voltage Scaling
LER	Line Edge Roughness
MODERN	MOdeling and DEsign of Reliable, process variation-aware Nanoelectronic devices, circuits and systems
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
NMOS	N-channel MOSFET
PCB	Printed Circuit Board
PDN	Power Distribution Network
PDK	Process Design Kit
PMOS	P-channel MOSFET
PoFF	Point of the First Failure
POST	Power-On Self-Test
PSS	Power Supply Selector
PVT	Process, Voltage and Temperature
PVTA	Process, Voltage, Temperature and Aging
RBB	Reverse Body Bias
RC	Resistor-Capacitor
RCA	Ripple-Carry Adder
RDD	Random Discrete Dopant
Si	Silicon
SiO_2	Silicon Dioxide
S-S	Slow-Slow
SOC	System on Chip
SSN	Simultaneous Switching Noise
TDC	Time-to-Digital Converter
T_{OX}	Oxide Thickness
TRC	Tunable Replica Circuit
ULV	Ultra-Low Voltage
ULV^2TC	Ultra-Low Voltage VTC
UTBB	Ultra Thin Body and Buried oxide
UPC	Universitat Politècnica de Catalunya

VCDL	Voltage Controlled Delay Line
VCO	Voltage Controlled Oscillator
VCTA	Via-Configurable Transistor Array
VLSI	Very-Large-Scale Integration
VTC	Voltage-to-Time Converter
V_{SB}	Source-to-Bulk Voltage
V_{TH}	Threshold Voltage
W2W	Wafer-to-Wafer
WID	Within-Die
ZBB	Zero Body Bias
ZTC	Zero Temperature Coefficient

PART I

GENERAL DISCUSSION

CHAPTER 1

INTRODUCTION & BACKGROUND

Technology scaling has made it possible to integrate more transistors into a single chip, increasing functionalities and decreasing the cost per unit. As a consequence, clock frequency has also been scaled up, so that these systems can compute faster more complex operations. In addition, supply voltage has been scaled down in order to reduce power consumption. However, as CMOS technology scales, PVTA variations have an increasing impact on performance and power consumption of electronic devices. Variability causes an undesirable dispersion of performance and power consumption and a consequent increase of chips per wafer which do not meet the expected specifications.

Yield is defined as the percentage of chips meeting specified timing and power constraints [23]. From an economical point of view, a 1-2% of Yield loss implies millions or even billions of dollars of revenue loss [24]. To tackle this issue, Design For Manufacturability techniques try to increase Yield by modifying designs making them simpler in order to make them more manufacturable, i.e., to improve Functional Yield, Parametric Yield, or reliability [25].

Adaptive techniques try to overcome the impact of PVTA variations by adjusting circuits once they have been manufactured. Post-silicon tuning is an effective solution to reduce the distribution of maximum clock frequency and power consumption, and thus, to improve Yield [26].

1.1 Trends of Scaling

In 1965, Gordon Moore predicted that the number of integrated transistors per die would increase by a factor of two per year, at least, until 1975 [27]. In 1975 Moore predicted a change in the slope based on economic reasons [28], with a doubling every two years, rather than every year. This law is still valid nowadays. In Figure 1.1 the increase in the number of transistors per die since 1971 can be observed. The channel length of transistors has been shrunk by a factor of more than 300 in 40 years. This has made it possible to operate at higher speed for the same power per unit area, since maximum operation frequency increases with transistor scaling.

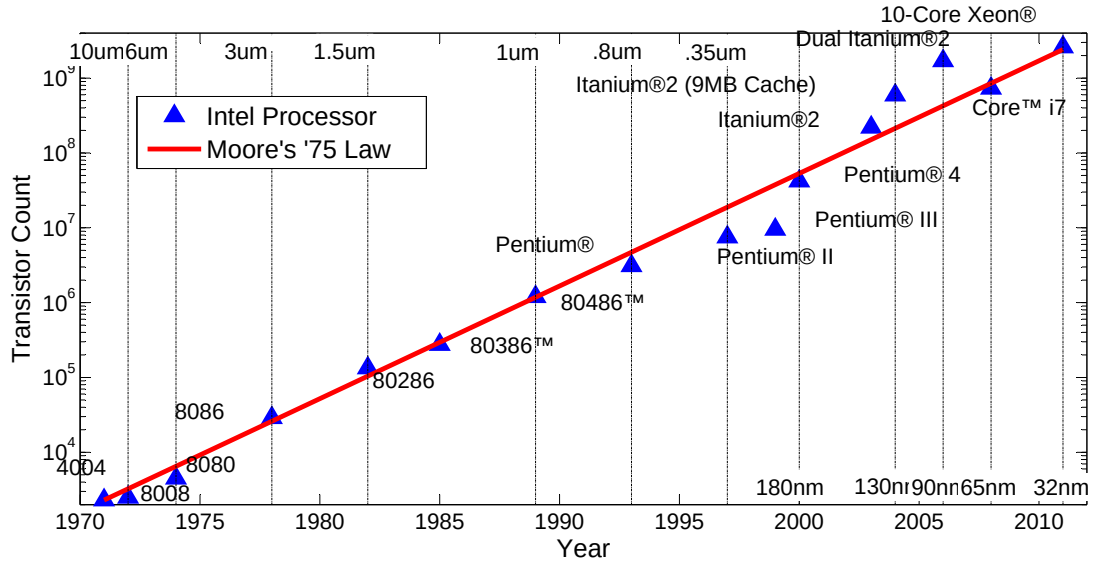

Figure 1.1 Moore's Law applied to the number of transistors per die compared to Intel Processors. Data from [1][2].

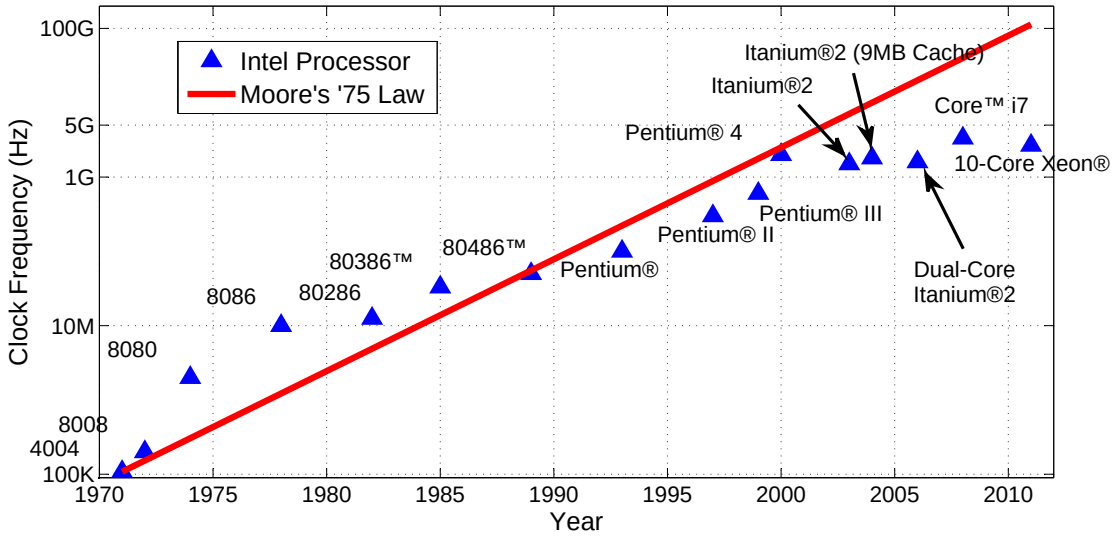

Figure 1.2 Moore's Law applied to clock frequency compared to Intel Processors. Data from [3][4].

Figure 1.2 shows that clock frequency also scaled up following Moore's law until the 2000s. However, it is clear to see the frequency hold back when magnitudes of GHz were achieved. Despite the fact that sub-micron transistors can switch faster, a number of reasons explain why frequency does not keep increasing with each generation. Firstly, parasitic capacitances induced by interconnections become critical in new technologies. Secondly, die area increases in order to make new processes economically viable, and thus, resulting in the increase of interconnection track length as well as its harmful effects. In fact, tracks have a big impact on metrics such as speed, power consumption and reliability [29]. Finally and more importantly, power consumption is directly proportional to clock frequency and power management and dissipation is one of the main problems with technology scaling.

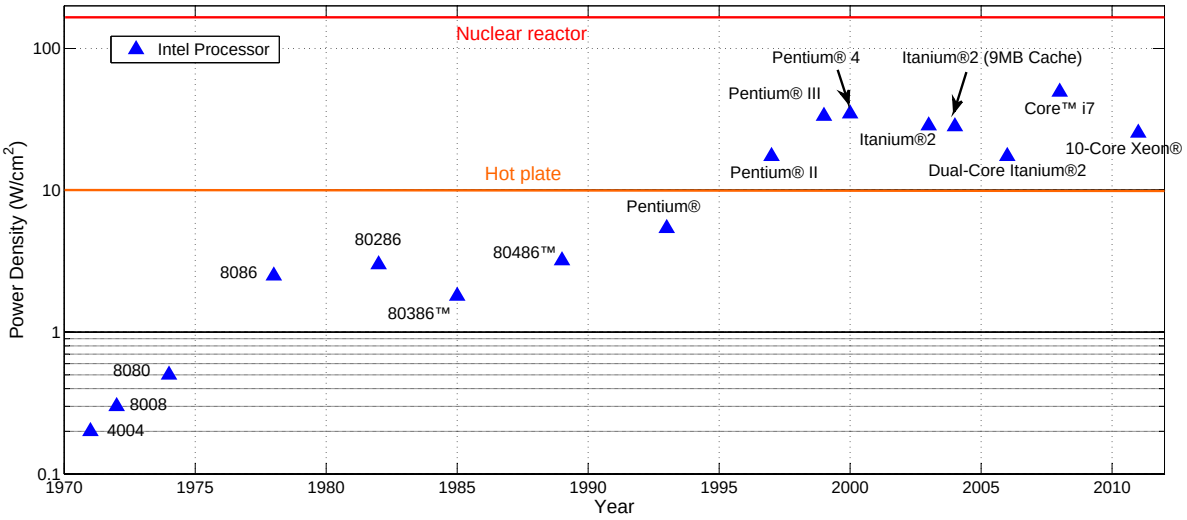

Figure 1.3 Power density in Intel Processors.

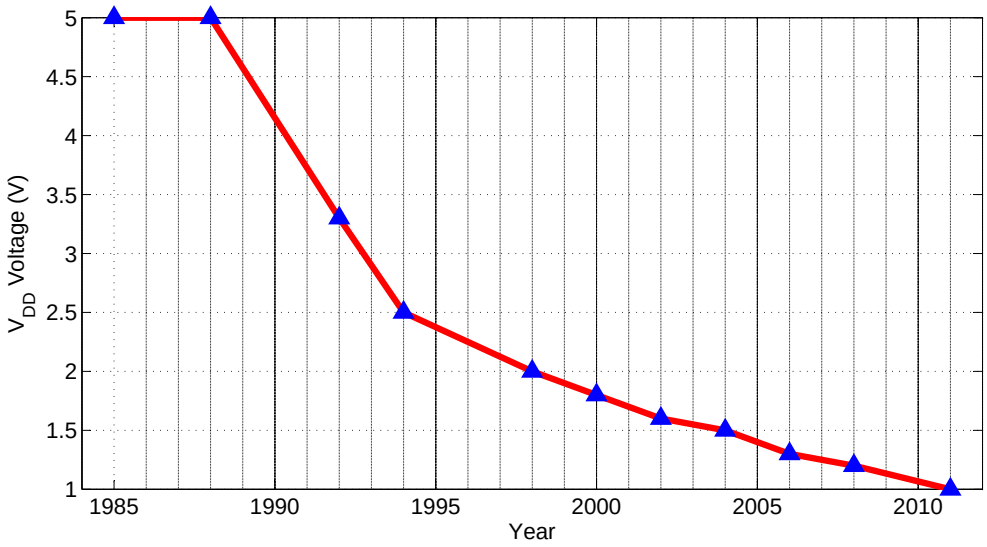

Figure 1.4 Evolution of V_{DD}.

According to Figure 1.3, power density has increased up to tenths of Watts per square centimeter. This can be attributed to the increase of clock speed observed in Figure 1.1. Both of them have a similar trend since power density has turned into one of the major upper bounds of clock scaling. Actually, temperature at the surface of an integrated circuit has a significant impact on the behavior, performance, and reliability of the semiconductor devices placed on it [30]. In order to reduce power density, and thus, power consumption, supply voltage has also been scaled. Figure 1.4 shows the evolution of V_{DD} during the last 30 years.

As technology scales transistor channel length, devices work in a velocity-saturated mode if supply voltage remains constant. In this scenario, a high power supply voltage implies a power penalty rather than improving transistors performance and, what is more, it may trigger reliability issues induced by hot-carrier effect and oxide breakdown [29]. For this reason, supply voltages have been reduced by a factor of five, and forecasts point to

further reductions. In [31], the author theoretically demonstrates that the minimum usable supply voltage for a CMOS inverter is about $8kT/q$, or 200 mV of power supply at room temperature.

However, scaling supply voltage while keeping threshold voltage constant results in an important loss of performance. Hence, threshold voltage has also to be scaled, causing an exponential increase in the device sub-threshold current as well [32]. This sub-threshold current, also known as leakage current, is said to be crucial in sub-micron ($<$250 nm) technologies, specifically for low power applications or portable devices.

In conclusion, CMOS technology scaling has made electronics more affordable to our society since the price per chip is continuously decreasing and, moreover, transistor scaling and SoC technologies have made it possible to integrate several functions into just one device, and thus answering the market needs. However, this continuous scaling has made circuits more prone to failure since error margins have also diminished, and hence, several physical and electrical phenomena no longer be neglected.

1.1.1 Challenges

According to the above considerations, transistor scaling has many benefits. However, PVTA variations arise as a result of transistor, supply voltage and operating clock frequency scaling. Variability produces substantial variations to physical characteristics of devices and interconnections. Supply voltage scaling, especially in near-threshold levels also makes variability of these physical parameters more important due to exponential dependence of currents in this regime. An increase of clock frequency, power density and temperature causes the appearance of temperature hotspots.

Consequently, electrical parameters of devices, such as threshold voltage, are affected and introduce variability in both space and time. Variability in electrical parameters is translated into variability in circuit characteristics such as leakage power, dynamic power and maximum operating clock frequency.

1.1.2 Design solutions

From the designer's point of view there are several reasons to continue improving designs or looking for solutions to reduce the problems that arise from technology scaling. The first and the most important reason is economical since the manufacturing cost of the newest CMOS process is prohibitive for some applications and for this reason a low yield implies a large economic loss. Another important reason is the fact that even new transistor generations such as FD-SOI and FinFETs are not exempt from PVTA variability and reliability issues. Undesired effects of CMOS transistor scaling can be tackled at two different design stages:

- **System/Block/Circuit level:** implementing sleep modes [33] or thermal monitoring [34] as well as the use of multiple voltage [35] or body bias [36] domains to manage power consumption and/or performance. Post-silicon tunable implementations give the possibility of avoiding or correcting malfunctioning issues, consequence of PVTA variations. In this thesis a novel circuit to reduce variability effects is proposed.

- **Cell/Transistor level:** the effects of variability can also be minimized by means of layout techniques. It is well known that regular layout structures suffer less from variability than non-regular layouts [9][37]. The main advantage of such techniques is that they do not require any post-silicon tuning mechanism, since variability is reduced by design. However, these kinds of implementations typically imply an important area penalty.

1.2 Types of Variability and its impact

PVTA variations have become a hot topic in the literature since the effects of these variations tend to be more harmful as technology scales down. Variability can be classified according to the source of variations, i.e. those static and dynamic factors which result in circuits not working properly, and according to the scale of variations, that is, variability between nearby transistors, across the silicon die, die-to-die or wafer-to-wafer.

Table 1.2 summarizes the main sources of PVTA variations and organizes them according to their spatial and time scale.

Proximity	Static	Dynamic	
		Reversible	Irreversible
Die-to-Die	Parameter means (L, T_{OX}, N_{SUB}...)	Operating temperature	BTI mean, HCI
Within-Die (systematic)	Lens aberration, Proximity effect, Voltage IR drops	Hotspots, SSN, Activity factor	Electromigration
Within-Die (random)	LER, RDD, IR	Self-heating	σ_{VT-BTI}

Table 1.1 Classification of Variations according to proximity and time dependence [22].

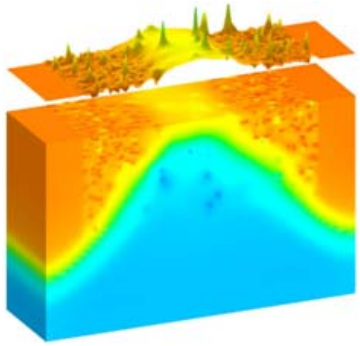

Figure 1.5 Electron concentration showing the impact of RDD [5]. Reprinted with permission.

1.2.1 Process Variations

Process Variations as its name indicates are those variations that are inferred to silicon chips during its manufacturing process. There are many factors involved in these variations, which are described below, causing reliability and power consumption issues to silicon devices.

1.2.1.1 Causes of Process Variations Process variations appear due to several different causes, some related to atomistic effects of the device structure and dimensions (Random Variations), and others related to manufacturing process imperfections (Systematic Variations). In sub-100 nm technologies, the number of dopants of each transistor channel is a relatively small number. As a matter of fact, the microscopic variations in the number and location of dopant atoms in the channel induce device RDD variations [38]. RDD is considered the most significant contributor to variability. In Figure 1.5 the electron concentration due to RDD variations can be observed.

Atomic-scale behavior of the manufacturing process creates missing chunks of atoms from the surface of the gate width, which is known as LER [23]. LER is produced by a statistical variation in the incident photon count during lithography exposure, and the absorption rate, chemical reactivity, and molecular composition of photoresist [22]. Figure 1.6 shows the potential distribution in device under the effects of LER.

The oxide layer used to separate the transistor gate from the substrate affects the electrical properties of the device. Oxide thickness (T_{OX}) is one of the limiting factors for device scaling due to the exponential relationship with gate tunnelling currents [39], which contribute to leakage power consumption. For this reason, the author in [40] suggests to scale down T_{OX} no further than 2 nm, which corresponds to 10 atomic layers. Variations in 1 or 2 atomic layers of T_{OX} cause significant V_{TH} variations that leads to leakage and performance variability [41]. This effect is known as Interface Roughness (IR).

Strain is also another cause of variability which has an impact on the electric resistance of silicon. Strain alters the band structure of Si, causing changes to properties such as bandgap, effective mass, mobility, diffusivity of

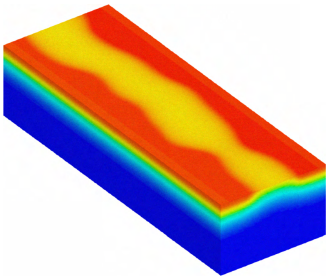

Figure 1.6 Potential distribution in a 30 nm x 200 nm MOSFET [6]. Copyright 2009 IEEE. Reprinted with permission.

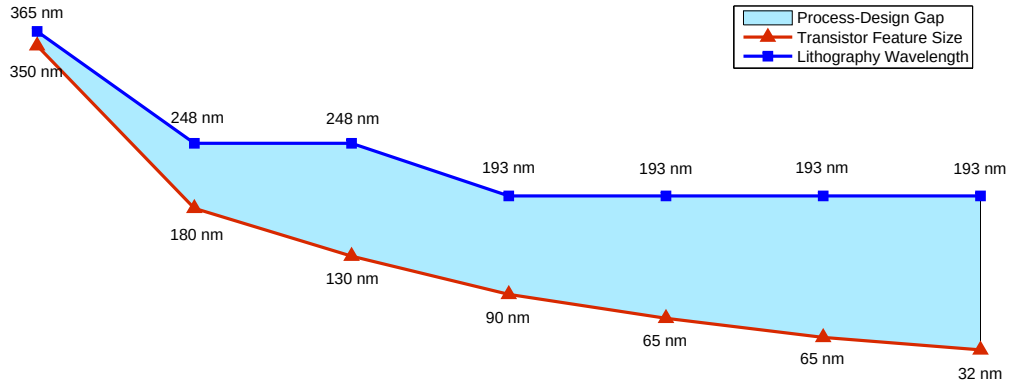

Figure 1.7 The gap between the wavelength of light used for lithography and transistor feature size [7].

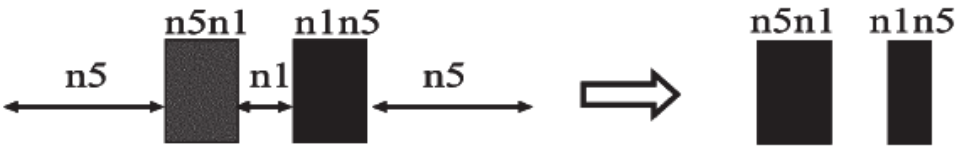

Figure 1.8 Coma effect. Two identical patterns are printed with different widths [8]. Copyright 2006 IEEE. Reprinted with permission.

dopants, and oxidation rates [42]. Regarding manufacturing process-related variations, we should mention possible inhomogeneity in physical and chemical conditions across wafer, and especially lithography related effects.

Current lithography tools have an illumination wavelength (193 nm) which is much greater than the smallest feature size IC manufacturers are attempting to print (see Figure 1.7. The increasing sub-wavelength gap causes very poor resolution and printability that, in turn, introduces substantial variability to the manufacturing and electrical performance yield of devices. Lens aberrations create optical path differences for each pair of rays through the imaging system. For example, Coma effect is a lens aberration that depends on neighborhood and location. It causes a gate surrounded by non-symmetrical structure to print differently from its mirror image [8].

1.2.1.2 Spatial Scale of Process Variations Process variations can be classified in terms of its spatial characteristics. They are classified in three groups: Within-Die, Die-to-Die and Wafer-to-Wafer [43]. WID and D2D classifications reflect some of the spatial characteristics of the variations along the wafer as shown in Figure 1.10. Those which vary slowly across the wafer are known as D2D (across-field) variations and they are caused by vari-

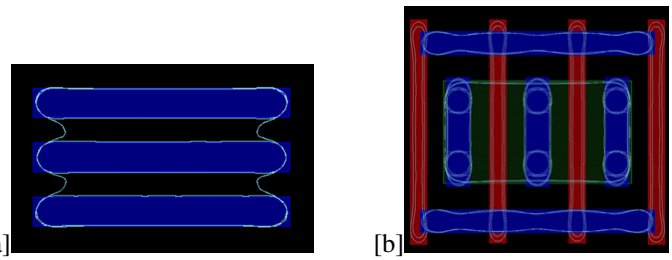

Figure 1.9 Proximity effect due to lithography [9]. Copyright 2010 American Scientific Publishers. Reprinted with permission.

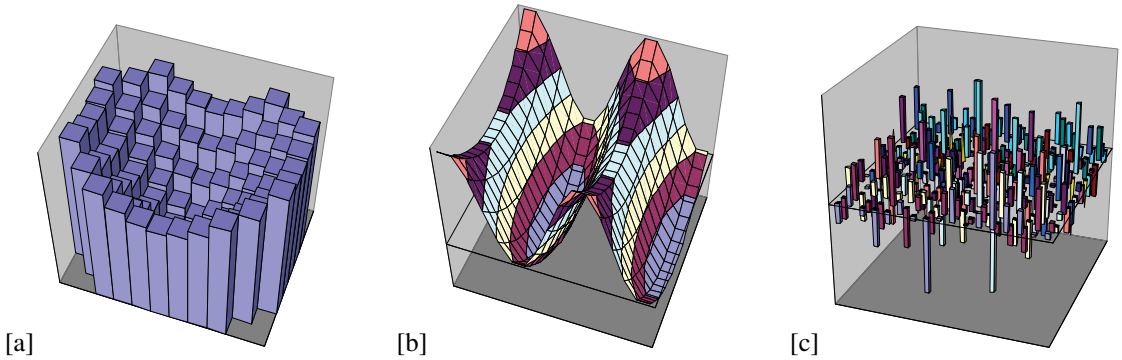

Figure 1.10 Scale of variatons: Die-to-Die [a], Within-Die systematic [b], Within-Die random [c] [10]. Reprinted with permission.

ations related to photo-litographic and etching process parameters. When variations change over distances smaller than the dimension of a die they are called WID (inter-field) variations.

In more detail, WID variations can be systematic or random. Systematic variations are typically induced by strain, lens aberration and layout dependent variations, among others. Random variations as a result of natural limits to scaling, and are mainly caused by RDD, LER and IR. Variations can cause differences in electrical characteristics of two identical devices with the same geometry, layout, and neighborhood [44].

1.2.1.3 Effects of Process Variations The effects of RDD and LER are changes on threshold voltage [45], which is a key factor of CMOS transistor performance and leakage power consumption. More precisely, leakage dramatically changes for small V_{TH} changes. IR effects also vary leakage but without modifying V_{TH}, that is, delay is not affected by IR [23].

From the spatial perspective, WID random variations individually produce changes to each transistor of the silicon die. This means a Δ_{Delay} that in relative terms will have a bigger influence in those gates with transistors with smaller channel width than in those with larger channel width. The reason lies in the fact that, according to Pelgrom's law, random variability varies as $\sqrt{WL}$. All these differences in delay are carried along circuit paths which results in overall path delay variability. What is more, the random nature of these variations increases the number of critical paths.

Power variability, and more precisely leakage, will suffer huge statistical deviations being more relevant in those transistors with smaller V_{TH}. The smooth shape of WID systematic and D2D variations produces an overall variation in power or delay in a certain die area (WID) or an entire die (D2D, W2W). This may cause the presence of hotspots as a consequence of systematic dynamic power increase.

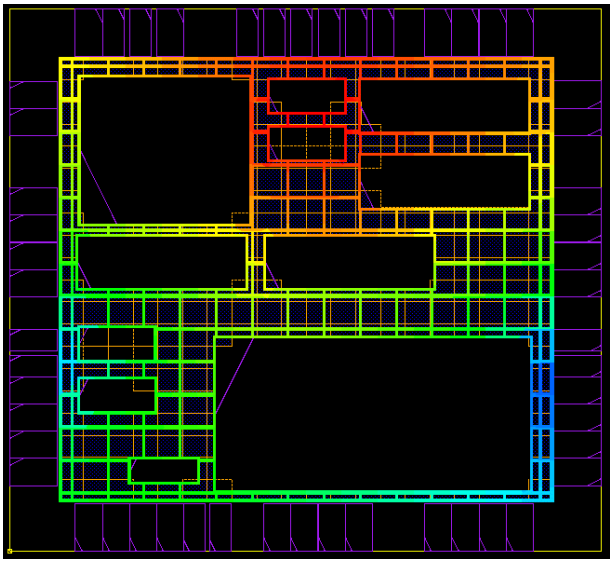

Figure 1.11 An example of an IR Drop simulation across power rails [11]. Copyright 2015 Teklatech. Reprinted with permission.

1.2.2 Voltage Variations

Power Distribution Networks also suffer from variability as a consequence of the continuous scaling of technology making systems more prone to suffer from noise.

1.2.2.1 Causes of Voltage Variations Two main factors can be distinguished: SSN and IR voltage drops. SSN is generated when a rapid switching of current occurs in inductive circuit components of the power/ground network. These inductive parasitic circuit elements of the network are due to the interconnection structures, such as the packaging, vias on the PCB, traces on the PCB, and decoupling capacitors. The SSN causes the voltage level at a position in a plane cavity to fluctuate; it propagates to other positions by electromagnetic propagation in the plane cavity and is also coupled to other plane cavities through cutouts or through other interconnection structures [46]. On the other hand, IR drops appear due to resistive losses across on-chip PDN [47]. Floorplanning and placement stages in complex designs with millions of cells are difficult when IR voltage drops are taken into account.

1.2.2.2 Spatial Scale of Voltage Variations Both SSN and voltage drops have a spatial dependency. SSN causes electromagnetic noise which is coupled through power rails interfering with nearby interconnections and transistors. From the cell or circuit perspective under a noisy environment, the closer and larger the source of SSN is, the higher the power supply fluctuations will be. IR voltage drops also depend on the physical arrangement of cells. If a power-consuming cell is placed far away from a power pad, or the supply current of a power network node is too large, the voltage drop becomes important.

1.2.2.3 Time Scale of Voltage Variations SSN temporal behavior depends on operation frequency and activity factor of circuits. More concisely, SSN can achieve resonant frequencies comparable with the clock frequency. For example, in [48], a 1 A triangle pulse of 2 ns period with rise/fall time of 300 ps is inserted into a circuit obtaining maximum resonant frequencies of the PDN at 260 MHz and achieving noise magnitudes larger than 30 mV. In contrast, IR voltage drops can be considered almost static since voltage fluctuations depend on the average current within a particular area. Hence, if the activity factor and thus the average current do not change significantly, changes in voltage drops are not appreciable with time. Switching between different power modes (e.g. from standby to active) or sudden changes in operating conditions may make these drops changing in time considerably.

Nevertheless, their frequency can be considered much lower than the clock operating frequency. This is especially critical since an IR voltage drop that persists over time can induce sustained timing errors.

1.2.2.4 Effects of Voltage Variations Voltage variations may impoverish the driving capability of logic gates and thus reducing circuit performance and reliability. While SSN may produce occasional clock violations due to fast decays on PDN, the effects of voltage drops last for many clock cycles. In the worst case, some parts of the silicon die may become unusable during a relatively long period of time. On the other hand, voltage surges produce an increase in dynamic power consumption (and thus thermal dissipation) that might lead to a hot spot generation.

1.2.3 Temperature Variations

Temperature Variations also play an important role in the overall variability of a circuit. They can be intrinsic, i.e. heat is dissipated due to transistor switching, or extrinsic due to environmental factors. What is more, thermal issues are aggravated in those circuits with a non-uniform power distribution such as CPUs. This non-uniformity along silicon die leads to a non-uniform temperature distribution in which hot spots can be observed, and such a chip will be more prone to suffer reliability issues.

1.2.3.1 Causes of Temperature Variations On the one hand, current density along power rails has increased substantially as a consequence of maintaining or even increasing power consumption. The increase in power density in new technologies has led to the appearance of hotspots and a decrease in reliability. On the other hand, changes in environmental temperature have to be considered. Silicon devices must work in an extended temperature range (of about hundred degrees or even more).

1.2.3.2 Spatial Scale of Temperature Variations Thermal distribution along silicon die can be decoupled into two components. The first one is an offset temperature which basically depends on the environment and the cooling system of the silicon die: packaging and heat sink -if any-. The second component depends on the activity factor of circuits, and hence, on the physical arrangement of components. A 2D thermal map will exhibit high peaks in those silicon die areas where maximum activity is produced since temperature dissipation is proportional to the power consumption. Therefore, the wider the high performance area, the wider the hotspot will be.

1.2.3.3 Time Scale of Temperature Variations Temperature can be static or dynamic. Static temperature is observed in those devices with constant power consumption, non-time dependent boundary conditions, and which have reached their thermal steady state. Otherwise, temperature depends on time and is dynamic. If so, thermal heat transfer function can be modeled as an RC mesh network with a low pass filter behaviur. Assuming that there is a direct proportionality between power consumption and temperature, high frequency components of the dissipated power will not produce any observable temperature increase due to the limited bandwidth of the thermal coupling. On the most optimistic scenario, temperature variations have a bandwidth close to 1 MHz [30].

1.2.3.4 Effects of Temperature Variations MOS characteristics like mobility and others are dependent on temperature. For this reason temperature variations imply variations in the electrical behaviour of gates.

In addition, temperature affects device reliability. Both Positive (for NMOS) and Negative (for PMOS) Bias Temperature Instability (BTI) has emerged as the primary MOSFET degradation and failure mechanism since the advanced sub-65 nm VLSI technology [49][50]. This phenomenon causes a gradual increase in V_{TH} of MOSFET devices that leads to reduced drain current causing larger delay of the cell, and to an eventual violation of timing constraints. BTI is severely worsened by temperature.

Self-heating due to the poor conductivity, affects channel current of the device through mobility, threshold voltage and velocity saturation mechanisms, worsening exponentially NBTI effects. The combination of these effects typically results in an overall reduction of drain current with the device temperature increasing [51]. As a consequence, temperature variations will influence the normal operation of circuit.

1.2.4 Ageing

Ageing effects in chips are also considered. Traditionally, ageing prediction was conducted under constant supply voltage, temperature and activity factor. However, as the degradation rate depends on the IC operating conditions such as temperature and input vectors [52][53], ageing can be considered to be a random process and its effects cannot be computed deterministically.

1.2.4.1 Causes of Ageing The predominant cause of Ageing is Hot-Carrier-Induced (HCI) degradation which occurs when electrons travel along transistor channels at high speed, shattering the surrounding silicon atoms once they have crossed the channel. BTI, already mentioned as a temperature effect, is also a very important cause of ageing. Electromigration, that is, the transport of material caused by the gradual movement of the ions in a conductor, also can be understood as ageing since this degradation increases with usage.

1.2.4.2 Time Scale of Ageing The common point of Ageing sources previously presented is their time dependent component. This effect, as its name indicates, is only appreciable after a long period of device usage. For example, over a period of 10 years running under nominal conditions, the V_{TH} of a PMOS device can increase up to 50 mV, causing timing violations [54]. Two identical chips with similar process variations and under similar environmental conditions may exhibit different performance and reliability if the difference of usage between them is significant (hundreds or thousands of hours). However, ageing has a strong dependency on non-predictable phenomena, such as supply voltage or operating temperature making the errors associated to ageing essentially unpredictable [52].

1.2.4.3 Effects of Ageing Gradual degradation as a consequence of ageing triggers reliability and performance issues. HCI degradation increases V_{TH} of devices making them gradually slower, and consequently path violations arise. The transport of material produced by electromigration forms voids on some parts of the interconnection, which produce the eventual or even the permanent loss of connections, while the acumulation of metal atoms will produce short circuits between adjacent interconnections [55]. Despite electromigration mainly affects those power rails with high current density, it also can appear on signal interconnections.

1.2.5 Impact of Variability on Parametric Yield in digital circuits

The effects explained above basically reduce yield and reliability. Yield is the probability that an integrated circuit meets power and performance constraints after being manufactured. Yield can be expressed as:

$$\boldsymbol{Yield} = \boldsymbol{P}(\max\left(\boldsymbol{Delay_{Path}}\right) \leq \frac{\mathbf{1}}{\boldsymbol{F_{CLK}}} \cap \sum \boldsymbol{Power_{Gates}} \leq \boldsymbol{Max_{Pow}}) \tag{1.1}$$

If all path delays are below clock period and the entire power consumption is also below desired target, a chip is considered good and can be sold. Obviously, the higher Parametric Yield achieved, the bigger revenues obtained.

Parametric Yield can be maximized in three ways: the first is by increasing tolerance margins, and thus reducing the amount of faulty chips. The second way is to implement DFM techniques, that is a set of different methodologies trying to make physical designs more lithography friendly to minimize the effect of process variations. The third way consist in the implementation of so-called post-silicon techniques. The aim of these techniques is to minimize the consequences of variability of a given IC once manufactured.

Increasing margins is no longer a desiderable solution due to the wide range of variations that chips may suffer. The challenge is to maximise Parametric Yield with the minimum possible tolerance margins as possible but also without increasing area overhead. When the number of manufactured chips is large enough, a simple 1% of Parametric Yield loss implies millions of dollars of revenue loss [24].

Parametric Yield maximization is one of the objects of study of this thesis. PVTA variations affect differently performance, leakage and dynamic power since they have different sensitivity to these variations. The study of correlations between the magnitudes listed before is important since high correlation between magnitudes implies that a cause-effect relationship can be established, and thus, one magnitude can be estimated from the measurements of this other magnitude.

1.3 Objectives of the thesis

The problem of PVTA variations and their effects on silicon devices has been studied for many years. Moreover, there are lots of proposals in literature to monitor and control the effects of variability, but the main problem of these proposals is their efficiency: some circuits provide a very fine grain variability control mechanism at the expense of large area penalty and thus power consumption. Other authors propose very simple and low area overhead circuits with a poor performance.

The main objective of this thesis is to design an efficient mechanism to reduce the effects of PVTA variations in CMOS digital circuits. To meet this requirement a set of more specific goals is listed below.

- **Study of variability sensors:** the first step to reduce the effects of variability consists in measuring a variable which indirectly points out the presence of variability. This objective goes further than studying which variables can be sensed: several types of variability sensors will be studied and the feasibility of their implementation in terms of area penalty will be assessed.
- **Study of variability mitigation techniques for short-channel technologies:** variability problems increase as transistor feature size diminishes. The objective is to study the existing variability mitigation mechanisms and understand its impact when applied together and separately. The objective is also to choose the parameter to be used for variability optimization. To do so, the effect of the parameter optimisation is studied by observing the effect in the non-optimized parameters and thus obtaining a picture of the overall variability in circuits after the parameter adjustment has been applied.
- **PVTA variability compensation circuit proposal:** the last objective in this thesis consists in the implementation of a circuit able to partially compensate the effects of variability. From the previously obtained results the observation variable is chosen and a circuit capable of sensing this variable is implemented. Finally, a generator able to influence over the circuit behavior and thus partially compensate variability.

1.4 Document Structure

This PhD dissertation is organized as follows.

Chapter 2 is devoted to the state-of-the-art of monitoring and control circuits. Firstly, it describes a previous work where a real-time PVT sensor is implemented, as a representative example of how PVT sensors can be implemented. Secondly, several Body Bias Generator (BBG) proposals are detailed. Thirdly, a study shows the effectiveness of post-silicon techniques based on electrical tuning of the circuit. Lastly, other post-silicon techniques (non consisting on Body Bias) are outlined.

A detailed analysis of variability indicators is presented in Chapter 3. Correlations between transistor device parameter variations, power consumption and circuit delay are studied. The objective of the study is to determine which sensing strategy gives the best picture about the overall variability in the chip. On the other hand, the feasibility of implementation of power and delay sensors is assessed.

Post-silicon tuning techniques are studied in Chapter 4. More precisely, it studies the effectiveness of Adaptive Body Biasing (ABB) and Adaptive Voltage Scaling (AVS). It also proposes a novel study to evaluate the impact of the overall variability reduction when different power consumption and delay are used as observable magnitudes. This information together with the obtained results in Chapter 3 is used to choose which variability indicator is used as the sensor of our variability mitigation circuit proposal.

The description and the measurements of the demonstrator chip prototype are shown in Chapter 5. The main objective of this demonstrator is to obtain real data about the spatial dependence of process variations (WID, D2D).

Chapter 6 proposes two BBG implementations to reduce the effects of PVTA variability in run-time, i.e. dynamically. The first BBG version is a very simple implementation with a very low area penalty but with low adjust range, while the second provides a wider adjust range. The main features of these proposals are the compactness, low power consumption, simplicity and scalability.

This thesis concludes in Chapter 7, which contains the conclusions derived from this thesis. Proposals of future work and research lines indicated by the research are suggested.

Notice that the full text of the publications carried out during this thesis is annexed at the end of the document with the aim of making this thesis dissertation easier to follow. The first article was presented in System-on-Chip Conference (SoCC) [56], the second was published in IEEE Transactions on Electron Devices (TED) journal [57] and the final article was presented in New Circuits And Systems Conference (NEWCAS) [58].

1.5 Methodology

In order to evaluate the proposals of the thesis, three main methods are considered:

- **Electrical simulations:** the main body of the proposals will be evaluated with electrical simulations based on schematic entry of selected circuits. Monte Carlo methods will be used to emulate variations in physical parameters. The available technologies in the group are 65 nm and 40 nm bulk CMOS and 28 nm FD-SOI, which will be the target technologies for the thesis. Computational capabilities of simulation servers as well as multithreading capabilities of simulation tools make it possible to run electrical simulations of complex circuits in a reasonable amount of time.
- **Physical design and electrical simulation of selected benchmarks:** when more realistic simulations are desired, extracted layout simulations may be considered. This implies the layout design of the schematic, Layout vs. Schematic (LVS) test and finally, extracted simulation. In all cases, simulation results are always analyzed from a critical point of view before accepting or rejecting hypotheses.
- **Silicon measurements:** whenever possible according to budget and time constraints, experiments on silicon are performed. Manufacturing services through CMP were used since this is the service offering 40 nm and 28 nm technologies. A chip in 40 nm was manufactured to test the effectiveness of ABB+AVS techniques.

CHAPTER 2

STATE OF THE ART

The most significant works on the state of the art related to the Thesis objectives are presented. Section 2.1 reviews monitoring circuits used for on-chip PVT sensing. Section 2.2 presents several works on BBG proposals. Works in Section 2.3 assess the effectiveness of ABB and AVS techniques as PVTA variability mitigators. Section 2.4 is devoted to other adaptive techniques (non-Body Bias) that try to minimize the impact of variability on CMOS circuits.

2.1 On-chip monitoring

Semiconductor device scaling has greatly improved performance and circuit integration density. However, integrated circuit performance and power density across a silicon die have become less predictable. This factor, in addition to multi-GHz operation of devices in which such critical paths are presented, causes the number of path delay faults observed on first silicon to increase significantly [59]. As a result, faults and performance limiters need to be identified as early as possible during design re-spin.

Thermal sensing has been widely reported in literature [60][61], as well as delay sensing [62][63][64].

For example, in [12] a fully PVT sensing strategy is presented in an Ultra-Low Voltage (ULV) application highly sensitive to Process, Voltage and Temperature variations. Thus, the author proposes a PVT sensor that provides real-time measurements.

The architecture proposed by this author is shown in Figure 2.1. The PVT sensor consists of four major blocks; they are temperature sensor, voltage sensor, voltage reference, and process sensor. Temperature sensor generates a frequency (f_T) proportional to the measured environmental temperature. The output frequency digital code connects to *clock* pin of the counter, and the counter can be positive edge triggered generating the corresponding digital output *T[9:0]*. The output *T[9:0]* represents the corresponding environmental temperature, but it is influenced by process and voltage variations. Therefore, a process sensor and voltage sensor are proposed to reduce PV variation of the temperature sensor. The real-time process and voltage information is also generated as *P[3:0]* environment and *V[4:0]*, respectively. The process sensor uses Zero Temperature Coefficient (ZTC) characteristic to design

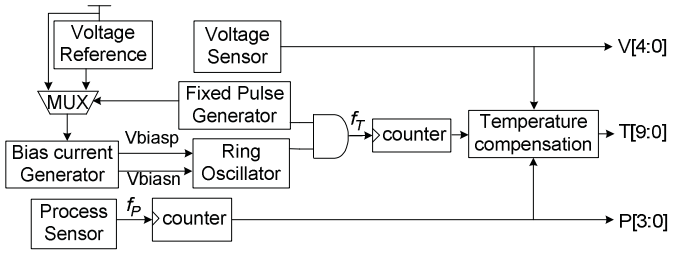

Figure 2.1 Block diagram of PVT sensor proposed in [12]. Copyright 2010 IEEE. Reprinted with permission.

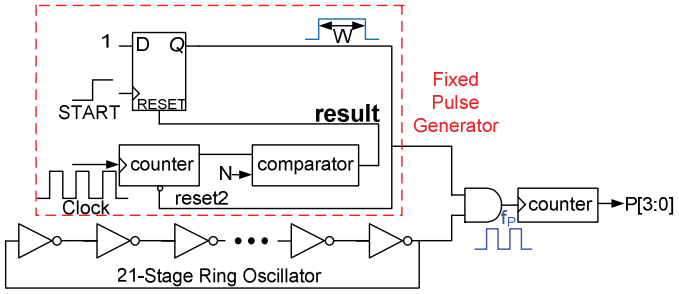

Figure 2.2 Process variations sensor proposed in [12]. Copyright 2010 IEEE. Reprinted with permission.

and it is independent of temperature variation. The process sensor also generates a frequency (f_P) proportional to the measured process corner. The operation is very similar to the temperature sensor. The voltage sensor uses a voltage-to-time converter, which measures inverter chain delay time of corresponding supply voltage. The *P[3:0]* and *V[4:0]* can not only reduce process and voltage variations of the temperature sensor, but also provide process and voltage environment information. Subsections 2.1.1 to 2.1.3 describe these sensors in more detail.

2.1.1 Process monitoring

The process sensor work as follows (see Figure 2.2): the pulse generator proposed by the author generates a pulse signal the width of which is independent to PVT variations. The pulse generator is composed of D-type FF, counter and comparator. When *START* signal rises, over a delay time (T_{d1}), the output of D-type FF also rises. When *result* signal rises, over a delay time (T_{d2}), the output of D-type F will be reset to 0. As delay times T_{d1} and T_{d2} produced by the FF suffer similar PVT variations, the effect of these latencies is considered to be static. From the above description, the output pulse signal width (W) is invariant to PVT variation. Monte Carlo simulations determine that process corner is linear to the frequency of ring oscillator, and thus, the counter converts the output frequency (fP) to 4-bit digital code *P[3:0]* from S-S corner (0100) to F-F corner (1111), which give information about the process.

2.1.2 Voltage monitoring

In order to sense Voltage, conventional high-speed and high-resolution ADC is discarded due to its high power consumption (about 50 mV) and chip area (1.2 mm^2). One way to overcome the challenge of the low-power and low-voltage design is to process the signal in time-domain. In voltage-to-time converter (VTC), the input analog voltage is converted to time or phase information. Thus, digital VTC circuit was presented in [65]. The circuit can replace conventional analog ADCs, and it can reach low-power, low- voltage, and small area. However, the VTC is not accurate with PVT variation. Therefore, a novel ultra-low voltage VTC (ULV2TC) circuit is proposed by [12] to improve accuracy, shown in Figure 2.3. This Voltage sensor converts input voltage to 5-bit digital code

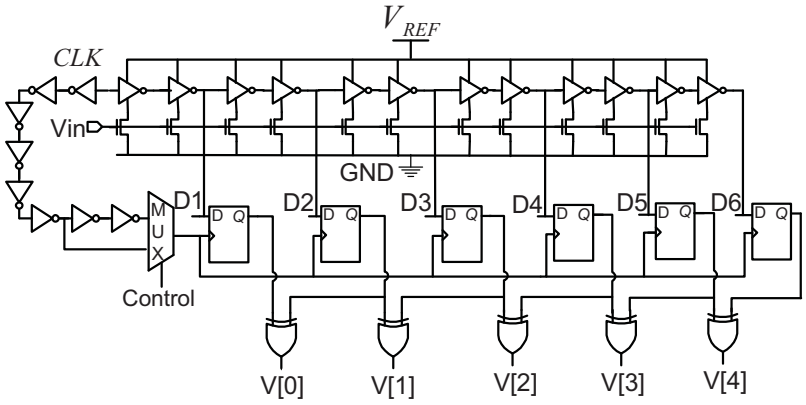

Figure 2.3 Voltage variations sensor proposed in [12]. Copyright 2010 IEEE. Reprinted with permission.

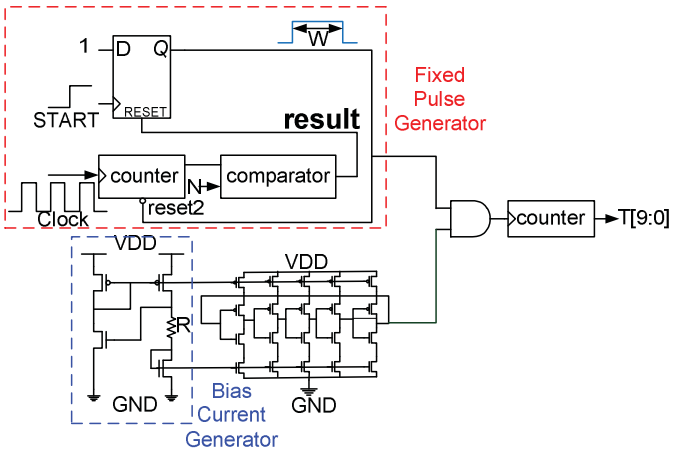

Figure 2.4 Voltage variations sensor proposed in [12]. Copyright 2010 IEEE. Reprinted with permission.

V[4:0] with a quantization step of 50 mV: from 0.3 V to 0.5 V (reference voltage). The PVT-aware ULV2TC consists of current starved inverters, FFs, and XOR gate. The control bit is from process monitor with the aim of compensating process variation, and thus, reducing the ULV2TC output error.

2.1.3 Temperature monitoring

Finally, the temperature sensing proposal of this work is shown in Figure 3.11. The output current of the bias current generator is proportional to temperature, which charges the inverters, so the frequency of the ring oscillator is also proportional to temperature with perfect linearity. The fixed pulse generator generates a pulse signal width independent of PVT variation. The ring oscillator output and fixed pulse through 2-AND gate, and the output frequency digital code connects to *clock* pin of counter, and the counter can be positive edge triggered generating the corresponding digital output *T[9:0]*. This output represents the corresponding environmental temperature. This value is adjusted according to PV sensor measurements in order to reduce the effects of process and voltage variations.

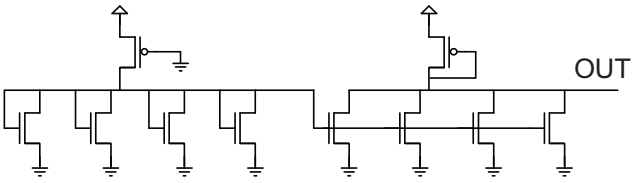

Figure 2.5 Monitor producing V_{SB} [13]. Copyright 2005 IEEE. Reprinted with permission.

2.2 Body Bias Generators

One of the main contributions of this Thesis is a BBG. A BBG is a circuit which is able to generate a bias voltage, which can modify the performance of a given circuit and thus counteract the effects of static and even dynamic variations once circuits have been produced (post-silicon tuning).

Body Bias technique consists in modifying threshold voltage (V_{TH}) in transistors by applying a source-bulk voltage (V_{SB}) different from 0. V_{TH} is the minimum gate-to-source voltage (V_{GS}) needed to create a conducting path between drain and source in MOSFET transistors. As can be observed from the equation 2.1, V_{TH} increases with V_{SB}. When a voltage higher than zero is applied V_{TH} increases and thus gate delay; this technique is called Reverse Body Bias (RBB). On the other hand when V_{SB} is lower than 0 gate delay diminishes at the expenses of leakage current; this technique is called Forward Body Bias (FBB).

$$\boldsymbol{V_{TH} = V_{T0} + \gamma(\sqrt{|V_{SB} + 2\phi_F|} - \sqrt{|2\phi_F|})} \tag{2.1}$$

$$\boldsymbol{CMOS_{Delay} \propto \frac{V_{DD}}{(V_{DD} - V_{TH})^{\alpha}}} \tag{2.2}$$

The major challenge in BBGs lays in the fact that RBB technique requires out-of-rail voltages (negative for NMOS and greater than V_{DD} for PMOS transistors). For this reason, many authors [13][15] only implement FBB which implies an important loss of variability adjust range. The second challenge is the number of Body Bias domains: short-range process fluctuations or local variations in voltage and temperature leading to a non-uniform behavior of the integrated circuit which cannot be recovered with a single Body Bias voltage. For this reason, several authors introduced the concept of Body Bias Islands (BBI) (also designated as partitions or domains) [66][14]. This concept consists in partitioning the circuit core into islands to control different regions of the chip with different bias voltages. This increase in granularity goes at the expense of area overhead: BBG circuits have to be replicated as many times as islands the chip contains.

In the following subsections some of the most relevant works in BBGs found in literature are shown.

2.2.1 Compensation for WID variations in dynamic logic

The authors in [13] propose a monitoring circuit that senses leakage variations and produces a V_{SB} voltage that compensates such variability. The adjust range is limited since it only implements FBB in the NMOS transistors.

The circuit works as follows: PMOS transistor in the left is always in saturation region since the gate is connected to ground. NMOS transistors on the left are configured as diodes (drain and source are connected together) in forward direction. As diode has losses drain voltages of transistors in the left will never get V_{DD} and therefore NMOS transistors in the right will be almost in the saturation region. As these transistors are not fully in the saturation region the output voltage will be slightly higher than 0 Volts. This operating point will vary according to the transistors V_{TH} and thus leakage variability is tracked. Reported V_{SB} is said to be 105 mV, which means that the circuit applies FBB by default.

The most interesting aspect of this paper is the low area overhead of this BBG. Despite the author does not provide any area dimension, this simple circuit is expected to be small in comparison with other proposals that integrates DACs to generate bias voltages.

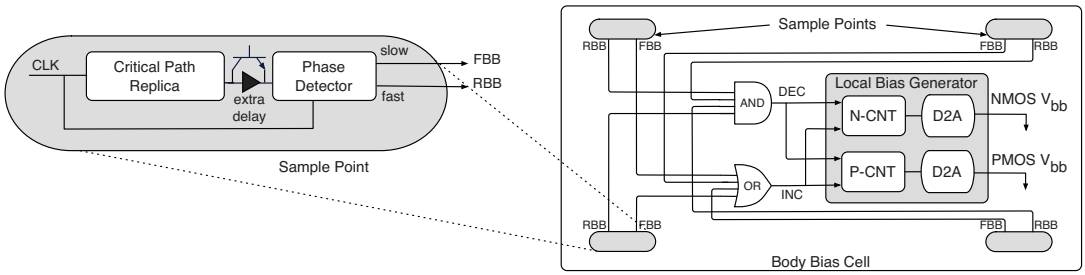

Figure 2.6 Fine-Grain Body Bias Generator circuit in a cell [14]. Copyright 2007 IEEE. Reprinted with permission.

2.2.2 Dynamic fine-grain body biasing

This paper [14] proposes a dynamic mechanism to compensate run-time variations by applying Fine-Grain Body Biasing (FGBB) with the goal of running a core at the highest frequency at the lowest possible power. The circuit is able to generate both FBB and RBB.

The author proposes to divide a chip into cells with independent Body Bias voltages. Each cell contains a BBG that determines the optimal Body Bias voltage according to multiple critical-path replicas distributed across the cell. Each critical-path replica contains its own phase detector, forming a Sample Point that outputs two flags: FBB and RBB. When phase detectors identify that critical-path replicas can go at higher frequency, a RBB flag is raised. When all the Sample Points in a cell enable a RBB flag, Local Bias Generator increases the NMOS counter and decreases PMOS counter. Oppositely, when at least one Sample Point detects that circuit performs worse than expected, NMOS counter increase and PMOS counter decrease. Finally, digital counter values are converted into a Body Bias voltage by means of two Digital-to-Analog Converters. These 5-bit DACs have -500 mV offset and 1 Volt of dynamic range so that output voltages are in the range between -500 mV for maximum RBB and 500 mV for maximum FBB in 32 mV steps. The effective V_{TH} adjust range is 70 mV. To deal with severe variation scenarios, an extra delay is added to one of the Sample Points in each cell. This extra delay is typically bypassed unless the cell fails meeting the target timing during testing.

2.2.3 Forward body bias generator with supply voltage scaling

In this paper a BBG with only FBB capability is proposed [15]. This $0.03mm^2$ block implemented in 90 nm low-power CMOS technology was designed to drive digital IP blocks of up to $1mm^2$. Similar to [14], the author generates both, PMOS and NMOS body biases by means of resistive DACs, but without RBB capability since output buffers are referenced to V_{DD} and V_{SS} (ground). The other significant difference is that the author does not consider the implementation of any variability sensing mechanism. Therefore, the circuit requires an external control unit to configure digital values corresponding to the body bias voltages to apply.

The circuit works as follows: the digital inputs $BBnw[5:0]$ and $BBpw[5:0]$, corresponding to PMOS and NMOS Body Bias configuration of the IP block, are decoded to generate row and column enable signals of the 8x8 polysilicon resistive array of the DAC. This DAC works in current mode, i.e. a constant current (I_{REF}) flow passes through an electronically tunable impedance (the resistive array) that produces a certain voltage drop. The maximum voltage drop across the resistor tree is 500 mV (maximum FBB of 500 mV), and hence the minimum step is 8 mV. I_{REF} is internally generated either by means of an external reference voltage (a bandgap circuit, for example) or an internal voltage (resistor tree as a voltage divider). Finally, an operational amplifier configured as voltage follower -with no gain- ensures low output impedance.

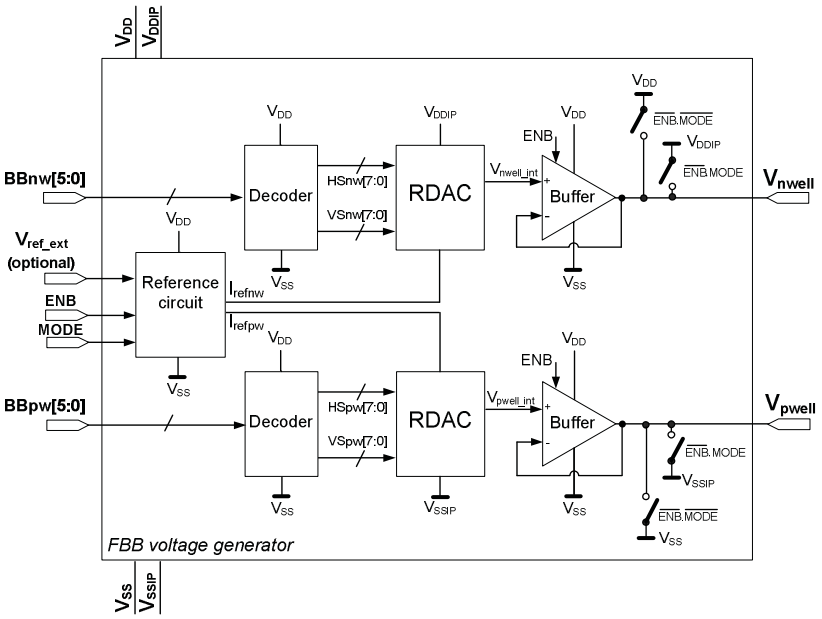

Figure 2.7 Block diagram of the BBG proposed in [15]. Copyright 2010 IEEE. Reprinted with permission.

2.2.4 Forward/reverse body bias generator for WID variability compensation

In these works [16][67][68] the authors propose a BBG with Forward and Reverse Body Bias capability. Despite the generator that the authors propose evolves with time, the essence of his proposal can be understood by looking at the block diagram shown in Figure 2.8.

Digital configuration inputs *A1[4:0]* and *A2[4:0]* are converted into analog voltages (*F1* and *F2* respectively) by means of a 5-bit R-2R lader type DAC, with 38 mV of resolution. To provide RBB capability, i.e. out-of-rail bias voltages, a Clock Generator block generates 4 clock signals (a pair of non-overlapping clocks *CK11*, *CK21* and inverted clocks *CK12*, *CK22*) from single clock and then these signals are coupled by means of transistors configured as capacitors (see Figure 2.9). Finally, a Charge Pump (Figure 2.10) converts DAC output into FBB or RBB voltage.

- **FBB Mode:** $\#EN$ signal goes low and Charge Pump circuit is disabled while *BP* goes high and thus *F1* is bypassed to *VNW*.
- **RBB Mode:** Charge Pump is enabled by setting $\#EN$ signal high. Then, DAC output signal *F1* is set to the left of the *M10* and *M20* coupling capacitors while *CK22* and *CK12* clock levels are low. After that, *CKP11* and *CKP21* go to saturation state and thus producing a capacitive coupling that brings *VNW* above V_{DD}.

2.3 Effectiveness of ABB and AVS

It has been widely reported by means of chip demonstrators that ABB and AVS techniques work well for power management purposes [69][70]. In [26] V_{DD} and V_{TH} scaling is taken on to increase yield. In other proposals, multiple ABB [36] or both ABB and AVS [71] voltage islands are proposed in multi-core processors to reduce power consumption or increase performance. In [72] a circuit to estimate V_{TH} and compensate it by controlling an on-chip ABB generator is presented. In this way, Process Variations can be reduced with a low area overhead since body bias voltage is directly generated from the V_{TH} sensor.

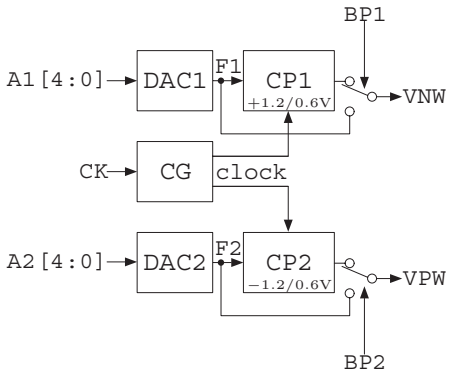

Figure 2.8 Block diagram of the BBG proposed by Kamae [16]. Copyright 2011 IEEE. Reprinted with permission.

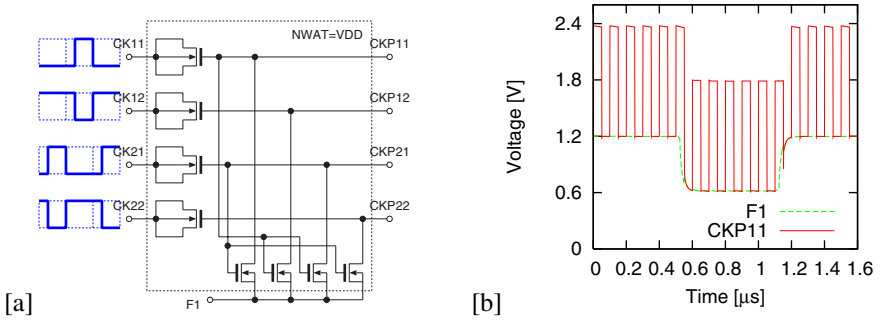

Figure 2.9 Over-V_{DD} phase clock generator schematic and simulation [16]. Copyright 2011 IEEE. Reprinted with permission.

In [17], results obtained from CMOS 90 nm technology show that performance and power consumption can be effectively adjusted by combining both techniques. In this study, supply voltage, P-well and N-well were scaled within the ranges presented in Table 2.3.

Operation	Swept Variable	Range (V)
AVS	V_{DD}	[0.6 , 1.2]
ABB	VN-Well	[V_{DD}-0.6 , V_{DD}+1.2]
	VP-Well	[-1.2 , +0.6]
ABB+AVS	VN-Well	[V_{DD}-0.6 , V_{DD}+1.2]
	VP-Well	[-1.2 , +0.6]
	V_{DD}	[0.6 , 1.2]

Table 2.1 Voltage ranges applied to power supply, P-well and N-well [17]. Copyright 2005 IEEE. Reprinted with permission.

A ring oscillator is used to assess the effects of ABB and AVS on performance and oscillation frequency when they are swept. Figures 2.11 and 2.12 respectively show the achieved frequency and power consumption for each V_{DD}, N-well and P-well combination. Red dots indicate those AVS configurations without body biasing, whereas blue cloud of dots depict the result of sweeping N-well and P-well given a fixed V_{DD} value. It is clearly shown that the combination of both techniques gives a wide degree of freedom in performance and power consumption according to circuit needs.

While AVS achieves a range of 24.4x power savings by 6.1x frequency downscaling from the highest V_{DD} to the lowest one, ABB+AVS yields 127x power savings with 37.4x frequency scaling from the highest V_{DD} and the

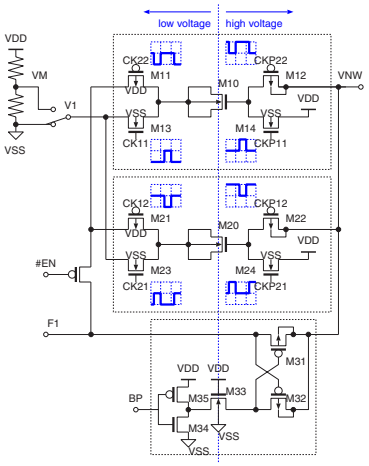

Figure 2.10 Charge Pump proposed by Kamae [16]. Copyright 2011 IEEE. Reprinted with permission.

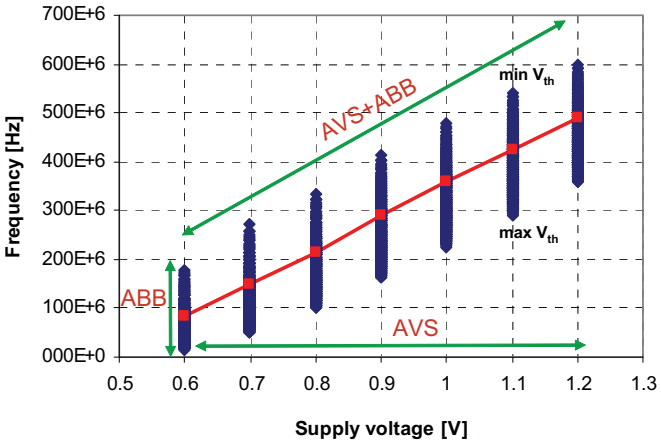

Figure 2.11 Oscillation frequency of the ring oscillator for each ABB, AVS configuration [17]. Copyright 2005 IEEE. Reprinted with permission.

lowest V_{TH} to the lowest V_{DD} and the highest V_{TH}. These results clearly justify the combined use of ABB and AVS.

2.4 Other Adaptive Techniques

2.4.1 TEAtime

TEAtime [18] uses tracking logic to mimic the worst-case delay in a synchronous system. Normally, the tracking logic is a one-bit-wide replica of the worst-case path in the system, with a slight delay added to it that provides a safety margin for the system.

The FF at the input to the tracking logic is wired as a toggle FF and clocked by the system clock, changing from 0 to 1 and 1 to 0 in alternate cycles. This provides a test signal for the tracking logic during every cycle of operation and ensures that the signal tests both types of transitions. The latter is necessary since delays for the two transitions can differ. The tracking logic output then goes through the safety margin delay. Next, the XOR gate

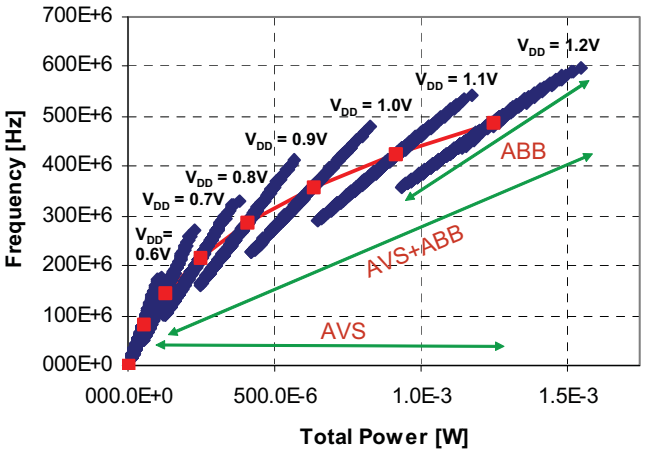

Figure 2.12 Frequency vs. total Power [17]. Copyright 2005 IEEE. Reprinted with permission.

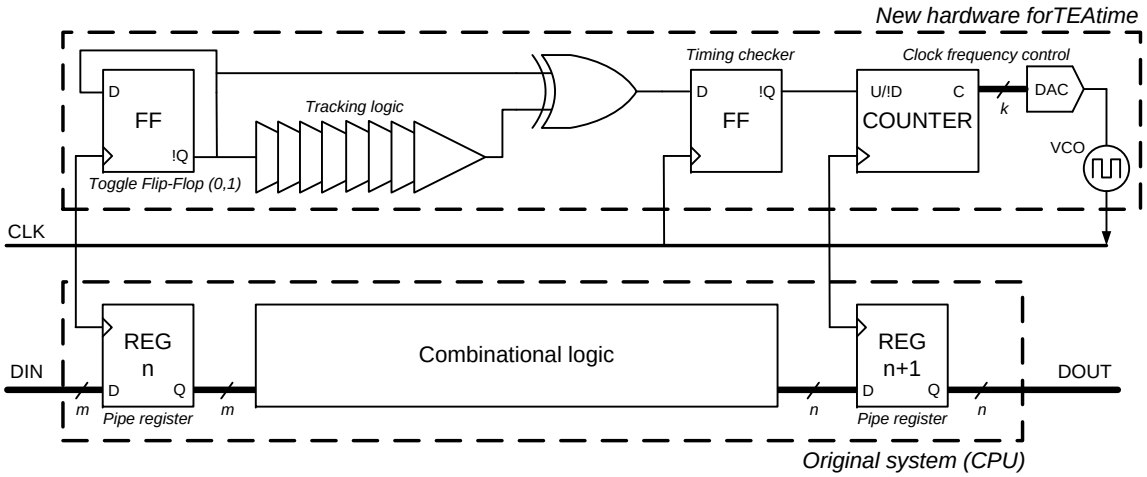

Figure 2.13 TEAtime design [18]. Note: m,n >> 1.

normalizes the test signal for the timing checker FF at the end of the chain. The timing checker FF also operates with the system clock. If the timing checker FF latches a 1, this means that the system clock period is close to being less than the worst-case path delay, and the system decreases the clock frequency. Conversely, if the FF latches a 0, the clock period is greater than the delay through the worst-case path of the real logic, and the system increases the clock frequency, improving performance.

The timing checker FF output therefore provides the command signal for the system clock generator to increase or decrease the clock frequency. This signal controls the counting direction of the up-down counter. The DAC converts the counter output to an analog voltage signal. This signal sets the clock frequency by controlling the Voltage Controlled Oscillator (VCO), which becomes the system clock. Thus, the clock period will never be less than the delay through the tracking logic plus the safety margin delay. Since the real logic of the system is as slow as or slower than the tracking logic, no timing error will occur in the real logic, which ensures correct system operation.

The weak point of this proposal is that it assumes uncorrelated short-range variations (mismatch) do not significantly affect the original system and that other static variability has similar impact on the original system and the tracking logic.

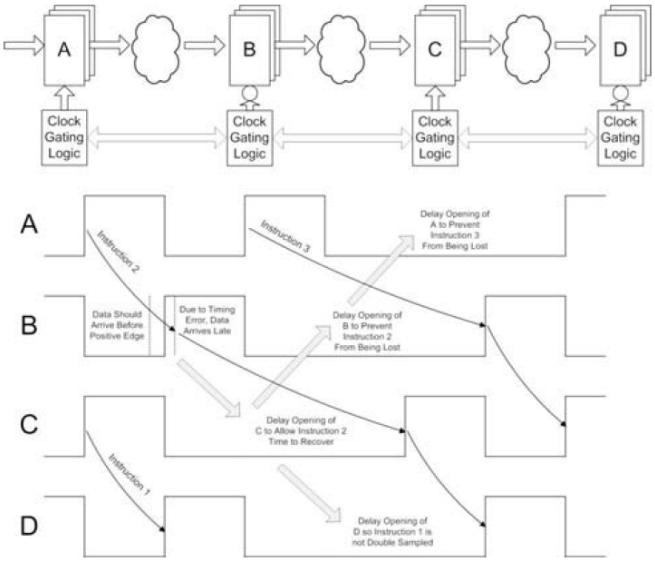

Figure 2.14 Error correction is done by propagating bubbles downstream [19]. Copyright 2012 IEEE. Reprinted with permission.

2.4.2 Bubble Razor

This method is a Razor-style system [73]. The aim of these kinds of systems is to eliminate timing margins by correcting transient path violations, using special FF. The novelty in Bubble Razor system is the non-invasiveness of the error detection technique (two-phase latches), that makes feasible to insert this system into a commercial processor [19].

This design allows timing speculation windows up to 100% of nominal delay, breaking the dependency between minimum delay and speculation window. This is specially attractive for those systems with large timing variations, such as low-voltage ones.

When an error occurs, B-Razor flags an error borrowing time from later pipeline stages. Error clock gating signals (bubbles) are propagated to each cycle from neighbor to neighbor, giving them one more cycle to get correct data. In order to prevent bubbles from being propagated along loops when circuit failures disappear, a novel bubble propagation algorithm is proposed (see also Figure 2.14):

- If a latch receives a bubble from one or more neighbors, it stops and sends bubbles to its neighbors (backward and forward), one half-cycle later.
- If a latch receives a bubble from all its neighbors, it stops but it does not send any bubble.

Bubble Razor was implemented in an ARM Cortex-M3 microprocessor replacing conventional FF by two-phase latches (3.29 times larger). At 85 C operating temperature, 10% of voltage drops, $2-\sigma$ process variation and 5% safety margin, the maximum operating frequency of conventional M3 processor is 200MHz. Tuning B-Razor performance and power consumption to the Point of the First Failure (PoFF), results are as follows:

Type of Die	**Slow**	**Average**	**Fast**
Maximum Frequency (MHz)	290	333	363
Power consumption @ 200 MHz	-43%	-54%	-60%

Table 2.2 B-Razor performance running at the PoFF [19]. Copyright 2012 IEEE. Reprinted with permission.

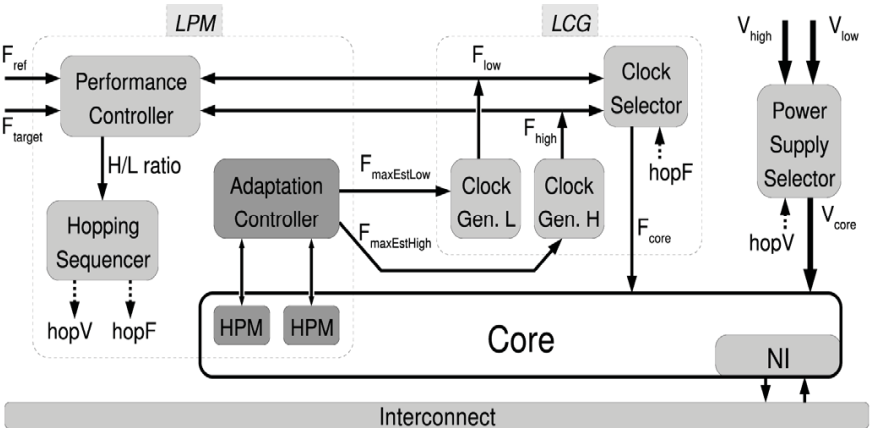

Figure 2.15 LAVS architecture proposed by CEA-LETI [20]. Copyright 2011 IEEE. Reprinted with permission.

Observing the results of the above table, power consumption is decreased when B-Razor runs at 200 MHz. However, this was achieved by manually reducing power supply to the PoFF, but this system does not implement any dynamic way to adjust supply voltage.

2.4.3 Local Adaptive Voltage Scaling architecture

CEA LETI has proposed a scheme to compensate variations, called Local Adaptive Voltage Scaling (LAVS) architecture (see Figure 2.15) [20]. In this system with independent voltage/frequency domains V_{DD}-Hopping is used, which means that fixed high-V_{DD} or low-V_{DD}, are used depending on whether high performance or low power consumption is required. For each voltage, mode high or low frequency mode is used. These frequencies are optimal taking into account the effects of variability for each domain.

This system works as follows:

- A Hardware Performance Monitor (HPM) is placed in the physical perimeter of the functional core and measures its performance variation due to variability, typically by means of delay lines of ring oscillators.
- Using the information from HPM, the adaptation controller estimates the maximum frequency the functional core can handle in F_{High} and F_{Low} modes with the minimum safety margin as possible. Low-jitter clock signals are generated at the selected frequencies.
- Using an external reference clock, Performance Controller controls high-V_{DD} and low-V_{DD} duty ratio in order to fulfill a performance target with the optimal power consumption. Hop sequencing is done by a Pulse-Width Modulator (PWM).
- Finally, Power Supply Selector (PSS), Clock Selector and the Sequencer make safe transitions and avoid clock violations in the functional core. To do so, the Sequencer must wait for PSS acknowledge signal before restarting the clock generation.

This alternative suffers the same problems as TEAtime. High Performance Modules only can estimate dynamic variations since the only way to evaluate the effects on static variations, and more concisely on random variations, is by using the core itself. The only way to ensure that no clock violations occur is by testing worst path delays and adjusting them.

2.4.4 Error-Detection Circuits

Intel has integrated resilient error-detection and recovery circuits in a 45-nm microprocessor with the aim of mitigate clock frequency guardbands needed to avoid clock violations due to dynamic variations, and thus increase

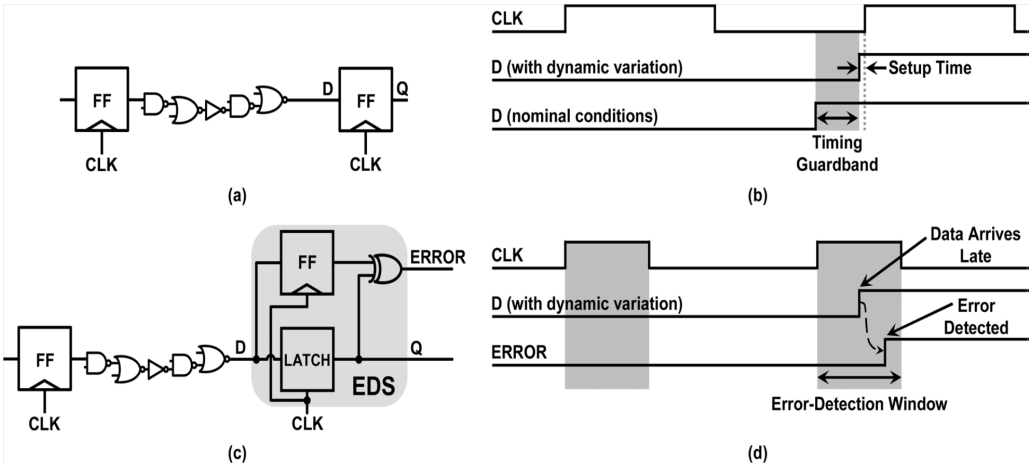

Figure 2.16 (a) Conventional design and its timing (b). (c) Resilient design by inserting EDS and its timing (d) [21]. Copyright 2011 IEEE. Reprinted with permission.

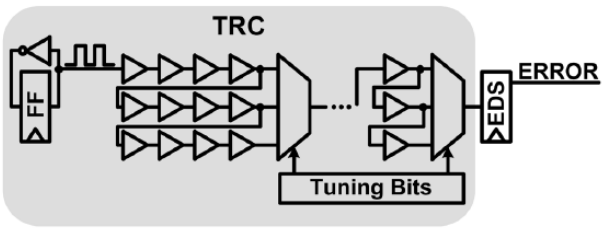

Figure 2.17 Tunable Replica Circuit [21]. Copyright 2011 IEEE. Reprinted with permission.

performance and efficiency [21]. They have implemented two separate designs: Embedded error-detection sequential (EDS) circuit and Tunable replica circuit (TRC).

The aim of EDS circuit is to detect timing errors on critical paths by adding time borrowing latch [74]. Figure 2.16 (a) represents the typical critical path. Figure 2.16 (b) represents the chronogram of this critical path. It can be observed that a timing guardband is provided so that worst-case dynamic variability effects cannot produce clock violations. In order to mitigate timing guardbands, EDS circuit is used instead of the receiving FF (see Figure 2.16 (c)). This circuit consists in two latches: the first one is a conventional FF which samples data during rising clock edge and drives data along datapath, while the second one samples data during high clock phase. If data arrives late, outputs of these latches will differ and error signal will be triggered. In Figure 2.16 (d) EDS timing is shown. Notice that timing guardband has been removed and clock signal duty cycle is lower than 50%. High clock phase length is equal to the error-detection window, that is, the maximum delay variation that EDS can detect.

TRC is a less-intrusive error-detection circuit which does not modify critical paths and it is based on a toggle FF and scan-configurable buffer delay chain (see Figure 2.17). A toggle FF switches the input every clock cycle and the output drives TRC, the path delay of which can be adjusted by setting multiplexer selection bits. The TRC output drives an EDS which will detect clock violations due to dynamic variations. TRC blocks are placed adjacent to each pipeline stage of the microprocessor. During the adjust time, the TRC delays of each stage are adjusted so that no timing violations occur on pipeline stages. Once TRCs are adjusted, the microprocessor starts working. Then, if dynamic variations cause a clock violation on the TRC, EDS triggers error signal. This method does not take into account critical path activations, i.e. EDS can generate an error signal despite no errors occurring on the pipeline.

The main advantage of EDS is that it can detect both local and dynamic variations while TRC cannot detect path-specific dynamic variations. However, it requires the adjustment of the clock duty-cycle and the minimum delay path in the core is restricted by the error-detection window. The main advantage of TRC is that it is a non-invasive method, and it reduces the design complexity overhead. Nevertheless, TRC requires delay guardbands to ensure that TRC delay is always higher than critical-path delays, and it suffers from false error-detection.

The main disadvantage of both proposals lays on the fact that logic must be redefined to be able to cope with errors, that is, data must be retained on all the stages of the pipeline when the error flag is triggered, otherwise data will be lost.

CHAPTER 3

VARIABILITY INDICATORS

The first contribution of this thesis is the study of variability indicators that let us understand the causes that produce variability and to what extent they affect to our circuits. Variability tests are done by means of Monte Carlo simulations. In order to check the feasibility of predicting variability of non-observed magnitudes from observed magnitudes, correlations between circuit performance distribution and static and dynamic power consumption distributions are studied.

Furthermore, the feasibility of several PVTA variability sensors is assessed. As for the leakage and dynamic power consumption sensor implementation, thermal sensing is considered given the expertise that the High Performance Integrated Circuits and Systems Design Group (HIPICS) from Universitat Politècnica de Catalunya (UPC) has in thermal sensing [30]. To implement circuit performance monitoring, two alternatives are considered: a delay sensor based in a Time-to-Digital Converter (TDC) and a delay sensor based in a Voltage Controlled Delay Line (VCDL).

3.1 Sensitivity of Power and Delay to PVT variations

The purpose of this study is to evaluate the impact of PVT variations to the static and dynamic power consumption and transistor delay assuming that only one parameter changes in the same way for each transistor. The obtained results give a clear picture of parametric sensitivity for each of the observed magnitudes.

Simulations were performed using BSIM4 device models for a 65 nm technology where the 3-σ variations considered are shown in Table 3.1 [75]. The testbench circuit used to perform these simulations is a chain of 15 inverters (see [56] in Appendix A).

Gate length variations are the most important source of leakage (see the exponential dependence in Figure 3.1). The second source of leakage is temperature, which in a very extreme temperature scenario can increase leakage power consumption by a factor of 3. Other parameters do not significantly affect leakage power consumption.

Parameter	L	W_P, W_N	T_{OX}	N_SUB}	V_DD}	Temp
3-σ var.	$\pm 20\%$	$\pm 5\%, \pm 8.5\%$	$\pm 10\%$	$\pm 10\%$	$\pm 10\%$	80 C $\pm 50\%$

Table 3.1 3-σ variability for each parameter

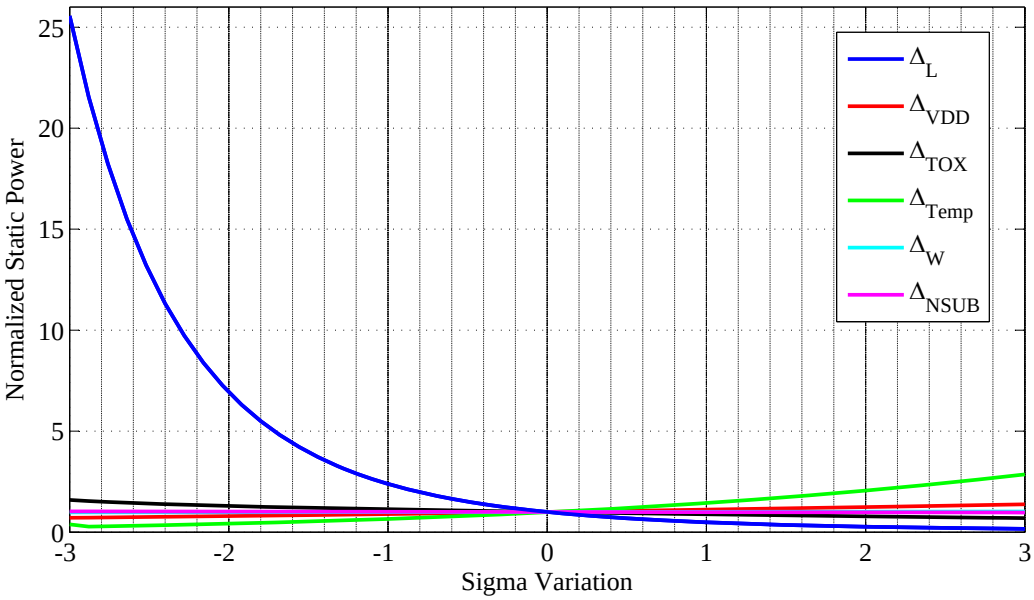

Figure 3.1 Effects of $\pm 3 - \sigma$ variations on normalized static power in a 15 inverters chain.

As seen in Figure 3.2, dynamic power mainly depends on supply voltage. Hence, if dynamic power sensor is used to detect variability, those variations produced by the power supply will be detected mainly. Dynamic power is not a good process variability indicator.

Figure 3.3 shows a quasi-linear dependence between the main source of process variations, i.e. channel length, and delay. Supply voltage also maintains a quasi-linear dependence with delay. Temperature variations do not produce significant changes on delay.

From the results above it can be concluded that transistor delay is very sensitive to transistor channel length and voltage variations, while channel dose and transistor width variations almost do not produce any observable change in delay. Delay variations produced by oxide thickness and temperatue are no further than $\pm$ 3%.

3.2 Statistical analysis of Power and Delay under PVT variations

The objective in this Section is to simulate the statistical effects of PVT variations in a circuit using the Monte Carlo method. Simulations were performed using BSIM4 device models for a 65 nm technology. The circuit under test is an 8-bit Ripple-Carry Adder (RCA) where power consumption and worst path delay are measured. In order to have sufficient statistics each simulation consists of 250 samples using a Gaussian distribution of parameters (3-σ) shown in Table 3.1. Results are represented using Cummulative Distribution Function (CDF).

The large static power variation shown in Figure 3.4 reflects the large influence of channel length fluctuations already seen in Figure 3.1. Around 25% of the circuits present more than twice the nominal leakage. This indicates that it is more difficult to bring leakage power consumption to its nominal value than to other magnitudes by acting on the circuit.

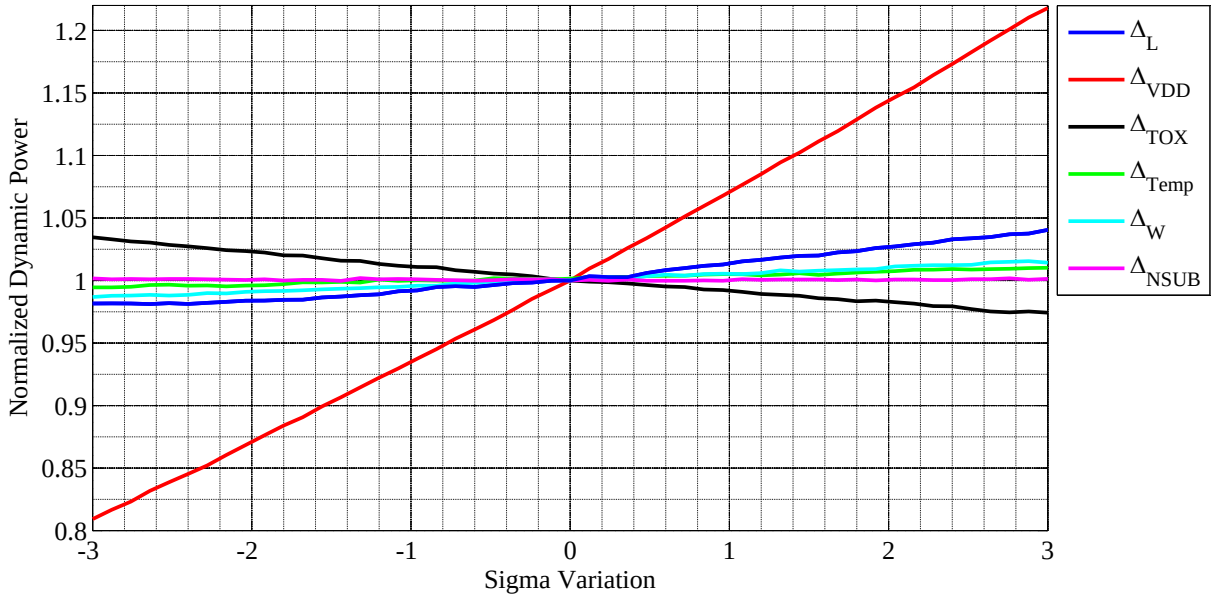

Figure 3.2 Effects of $\pm 3 - \sigma$ variations on normalized dynamic power in a 15 inverters chain.

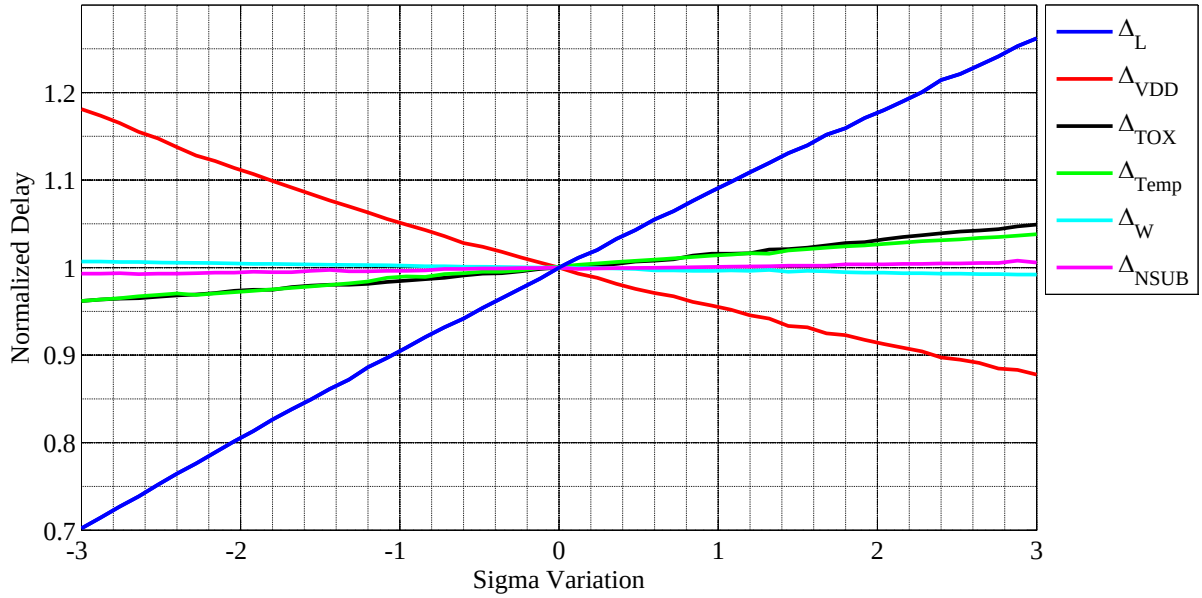

Figure 3.3 Effects of $\pm 3 - \sigma$ variations on normalized delay in a 15 inverters chain.

In Figure 3.5 and Figure 3.6 showing the distribution of dynamic power and delay, it can be seen that variations are smaller and both of them obey to a Gaussian distribution. This can be attributed to the quasi-linear dependence between these two magnitudes and PVT parameters which, in turn, follow a Gaussian distribution.

It can be concluded that static power consumption is highly variable due to the exponential dependence on transistor channel length. As for delay variability, around 20% of the Monte Carlo samples would not pass performance test even setting a 10% of Design Margin.

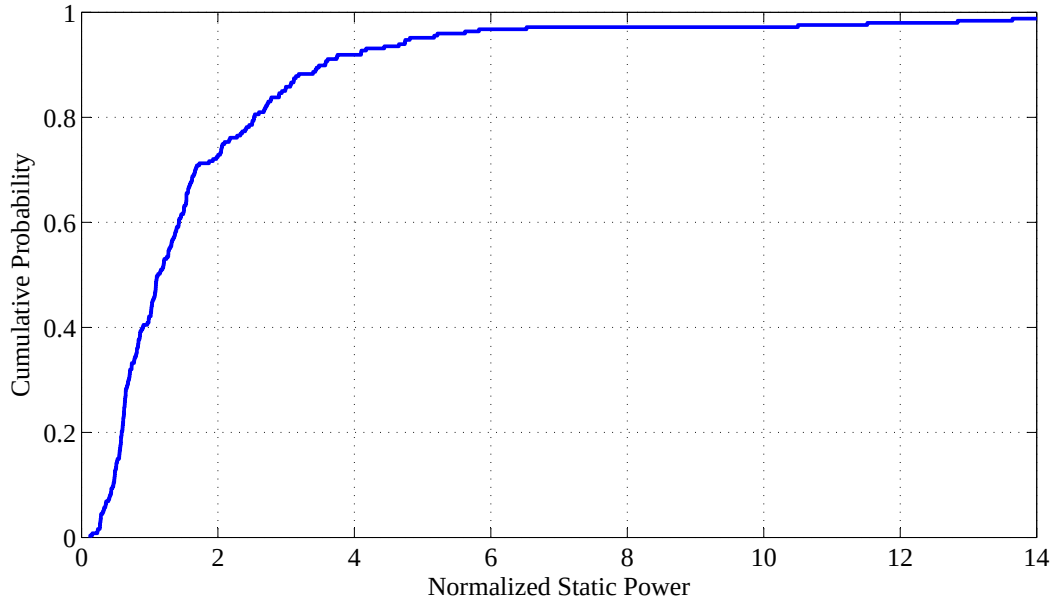

Figure 3.4 8-bit RCA normalized static power CDF.

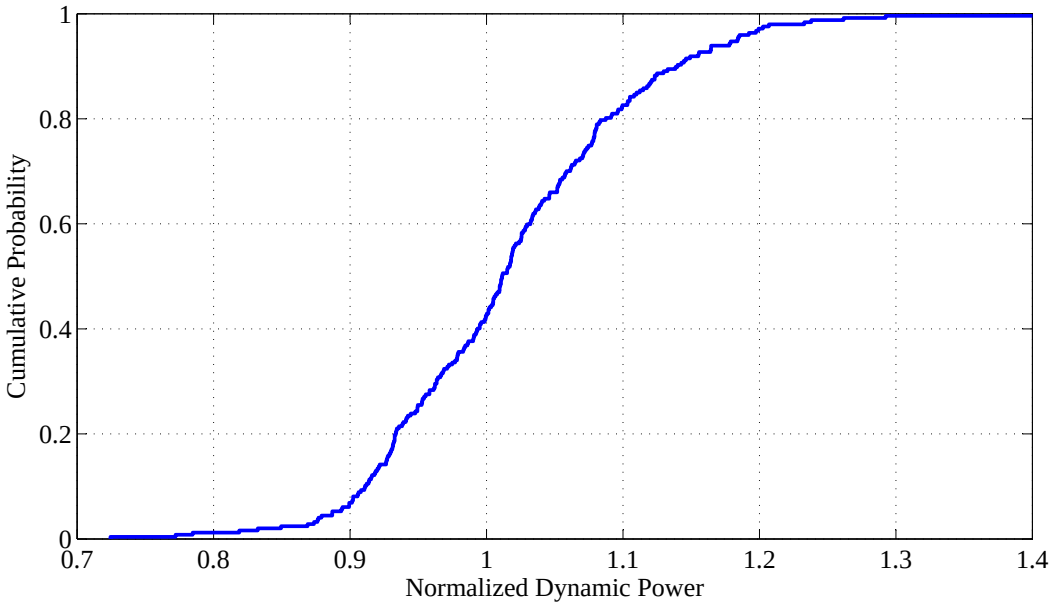

Figure 3.5 8-bit RCA normalized dynamic power CDF.

3.3 Statistical correlation between variability indicators

The correlation between power and delay magnitudes has to be taken into account in order to determine if non-observed magnitudes can be estimated from the observed magnitude. Control of variations is based on a sensor measuring a given magnitude (e.g. delay). Then, circuits which exhibit different characteristics to those expected (nominal) are compensated by using a given control mechanism and variability is thus reduced for the observed magnitude.

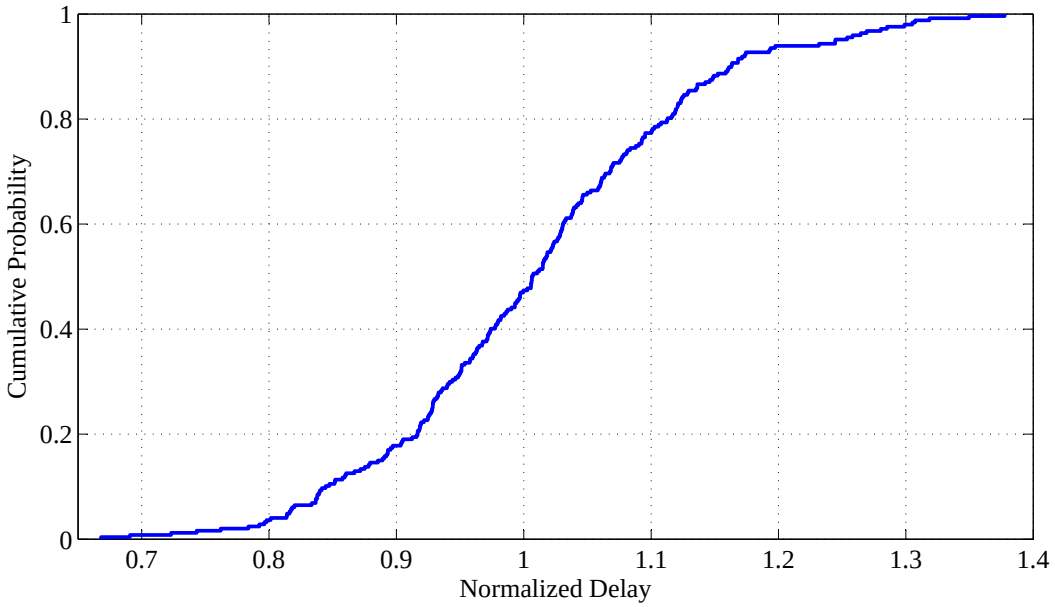

Figure 3.6 8-bit RCA normalized critical path delay CDF.

A high correlation between magnitudes implies that when using control mechanisms to reduce variability in the magnitude being measured, the variability in the correlated magnitude is also reduced. The higher the correlation between magnitudes is obtained, the higher clock frequency or power consumption variability reduction can be achieved when adaptive techniques are applied. Inversely, when no correlation exists between two magnitudes, the application of adaptive techniques based on the observed magnitude will increase the second magnitude variability, rather than bringing it to its nominal value.

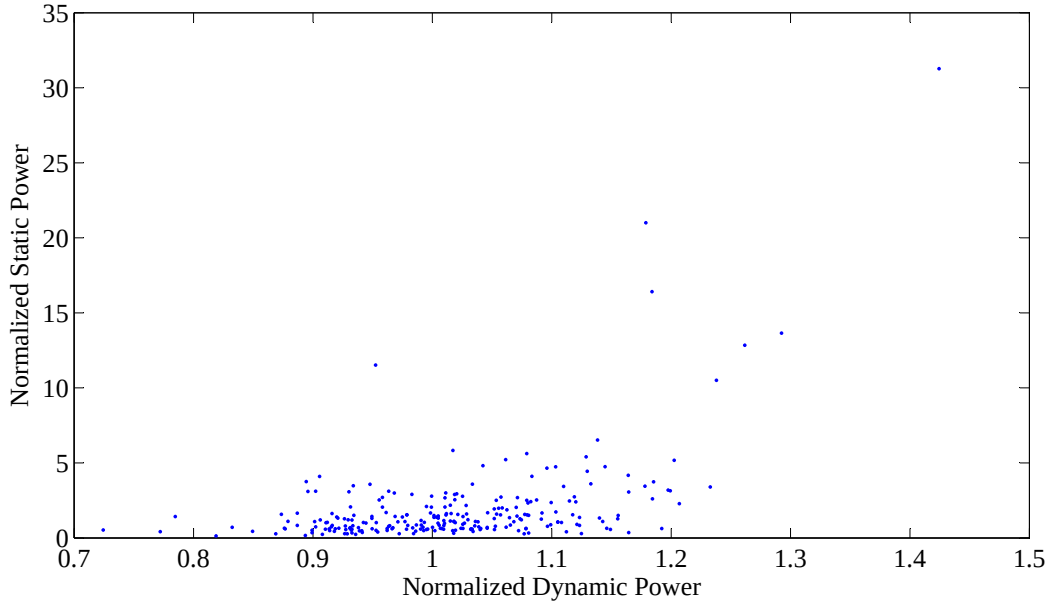

Figure 3.7 Normalized static power vs. normalized dynamic power scatter plot in an 8-Bit Ripple-Carry Adder. Correlation coefficient = 49%.

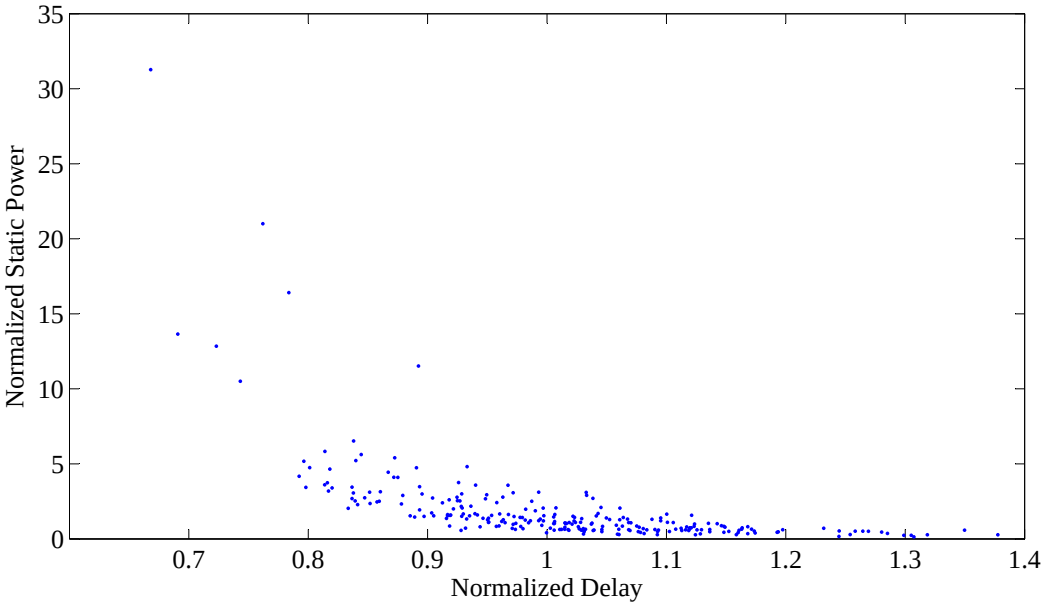

Figure 3.8 Normalized static power vs. normalized delay scatter plot in an 8-Bit Ripple-Carry Adder. Correlation coefficient = 62%.

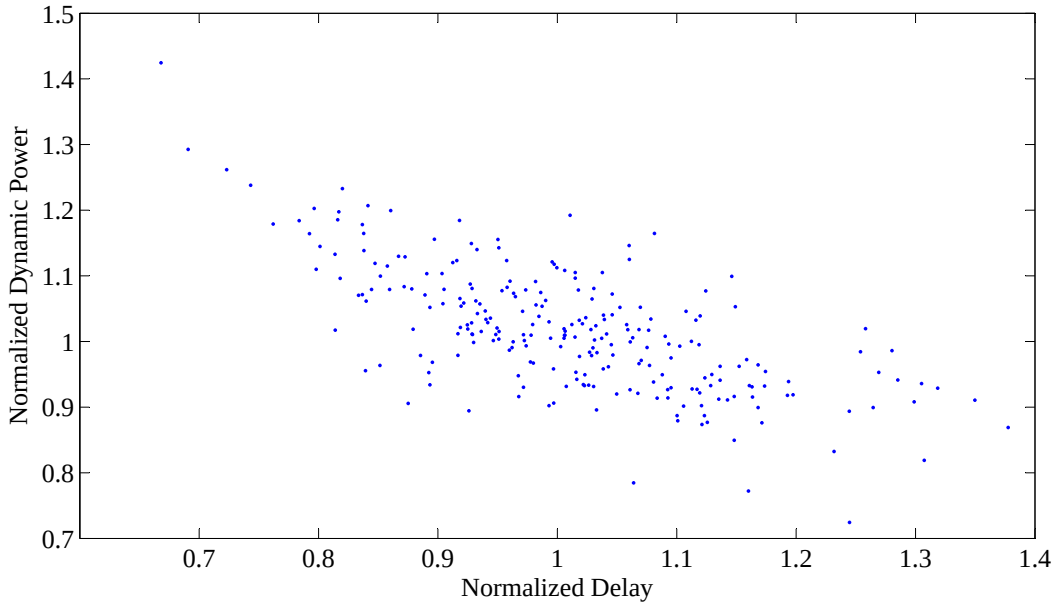

Figure 3.9 Normalized dynamic power vs. normalized delay scatter plot in an 8-Bit Ripple-Carry Adder. Correlation coefficient = 70%.

Figure 3.7 shows correlation between static and dynamic power. Correlation coefficient between both magnitudes is very low: 49%. This means that static power cannot be accurately predicted from dynamic power measurements, and vice-versa. This lack of correlation will cause an increase in dynamic power variability if some adaptive technique is used to decrease leakage power variability, and vice-versa.

The correlation coefficients between static power and delay (Figure 3.8) and dynamic power and delay (Figure 3.9) is 62% and 70% respectively. This means that delay variability suffered by transistors is better correlated with the other two power consumption magnitudes. Therefore, it is expected that the use of transistor delay to compensate variability not only will reduce circuit performance variability but also power consumption variability.

3.4 Thermal characterization of power consumption

The main objective of this study is to get an estimation of the temperature that transistors dissipate in their standby mode and during their maximum activity rate. In this way, it can be assessed the suitability of thermal sensors to monitor power variations, either Dynamic or due to Leakage.

This study is done by means of numerical methods, more concisely RC modeling of heat transfer [76]. This method consists in discretization of the region under analysis into a mesh of nodes and generates a set of linear equations where the unknown quantities are the temperatures of the different nodes. This model establishes an analogy between voltage and current (electrical domain), and temperature and power consumption (thermal domain) respectively.

Each transistor is considered an independent heat source whose dimension depends on the active region (the main heat source) of that transistor, i.e. WxL. This heat source is modeled as current source with a current equivalent to dissipated power. Thermal analysis is done assuming linearity; that is to say, heat sources are independent among them. Grid resolution as well as material properties will configure the RC network.

3.4.1 Dynamic power thermal characterization

A ring oscillator of 187 stages is considered for this test. The advantage of using such a kind of circuit is that dissipated power consumption remains very stable due to the fact that the number of switching transistors is always the same. This is beneficial for two reasons: the first is that power consumption measurements will remain very stable unless dynamic variations occur, and the second is that switching noise produced by ring oscillator and coupled to power supply rails is very low since current consumption is almost constant.

Physical placement of transistors in the circuit is crucial for the analysis. For this reason, an inverter cell from a 65 nm Process Design Kit (PDK) library was considered. As a first approach, the total area of the sensor is set to 15x15 μm^2 and assuming that 5x5 μm^2 would be enough for thermal sensor circuit, the number of inverter stages which fit this area are 187.

The electrical simulation of power consumption of each inverter stage is computed. Introducing this information as well as the physical location of each inverter into the RC model simulator, 2D power consumption map is obtained (see Figure 3.10). It can be observed that this 15x15 μm^2 ring oscillator produces about 4 ^oC in the middle region where the thermal sensor would be placed.

In conclusion, dynamic power sensing by means of temperature sensors is feasible, since the thermal gradient produced by few logic resources is in the order of degrees. Therefore, when placed into a physical design with thousands of gates around, the expected thermal gradient will be larger enough as to to be measured with good resolution. Notice that dissipated heat exponentially decreases with distance, which means that thermal sensing is spatially limited.

3.4.2 Leakage power thermal characterization

Taking into account that leakage power consumption is about four orders of magnitude below in comparison with maximum power consumption of a transistor, a chain of 15910 inverters is considered. In this case, the chain of inverters has the input stuck to ground so that transistors only dissipate static power.

Inverters cells are mapped to a 130x130 μm^2 square region where each cell is dissipating a certain amount of power consumption. With the physical coordinates and cell power consumption 2D thermal map is computed by the RC model simulator. According to Figure 3.11, this circuit produces around 4 m^oC in the place where maximum temperature is achieved.

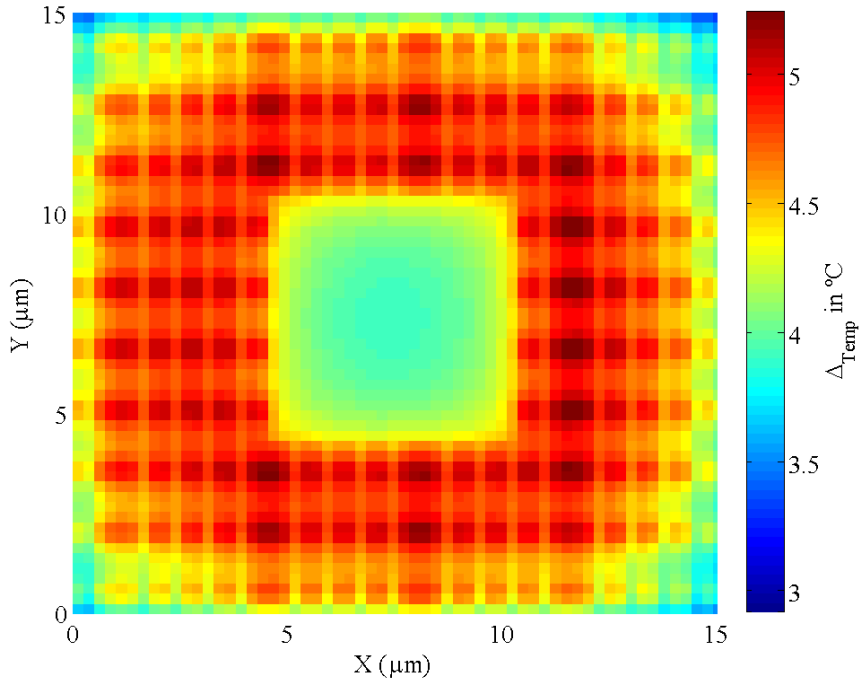

Figure 3.10 Dynamic power thermal map of a 187-stage ring oscillator.

Leakage thermal sensing is not a practical approach since the heat dissipated by inverters in standby mode is not big enough to be sensed with a good resolution. Dissipated heat by transistors exponentially decreases with distance and hence dynamic range of leakage thermal coupling would not increase substantially even placing more inverters. Anyhow, this would imply a huge temperature sensor and, what is more, temperature measurement could easily be masked by further active heat sources which operate during leakage power sensing, e.g. leakage control circuit itself.

3.5 Circuit delay sensor alternatives

As seen in Section 3.3 delay of transistors is a good variability indicator. For this reason, two alternatives based on sensing a delay line are proposed in this Section: a TDC-based delay sensor and a VCDL-based delay sensor. Working in the digital domain has two main advantages: a digital sensor can easily be reconfigured and sensor status can be externally monitored without any extra hardware. However, at some point the obtained digital word by this sensor will need to be converted to analog domain in order to adjust any of the tunable parameters of the circuit thus causing a large area overhead. This fact makes a VCDL-based alternative necessary whose main feature is its low area penalty.

3.5.1 TDC-based delay sensor

The first sensing approach to be considered is based on a TDC which has a simple operating principle (see Figure 3.12). A toggle FF switches the input of an M-stage delay line. The output of the XOR array in Figure 3.12 produces a thermometer code indicating the number of delay line stages that the launching signal can propagate within one clock period. Then, using a priority encoder, the delay of the chain can be quantified with $N = log_2 M$ bits. Finally, measured delay is compared with a reference delay value and the output of this comparison can be used to increase or decrease some control magnitude, such as V_{DD} or V_{TH}.

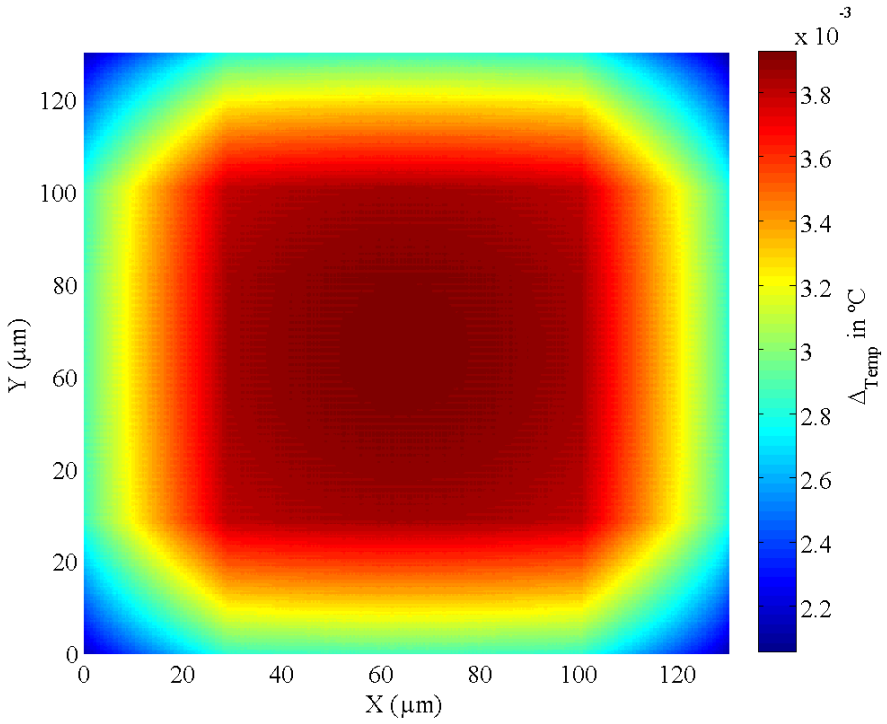

Figure 3.11 Static power thermal map of a 15910-stage inverter chain.

Simulation results in Figure 3.13 show the TDC-based delay sensor working in open loop, i.e. without using comparator output signals to increase or decrease V_{DD} nor V_{TH}. In order to test the effectiveness of the sensor, dynamic variations were induced to supply voltage by ramping it up from 0.6 to 1.5 volts, and then ramping it down to 0.6 volts. $INST\_SPEED$ and CAL_SPEED are 6-bit vectors that represent the actual circuit speed and the measured speed after calibration period. Calibration is very useful to compensate non-correlated process variability (mismatch). It can be seen how $INST\_SPEED$ tracks power supply variations, and how *INC* and *DEC* flags rise when performance is above or below the threshold.

3.5.2 VCDL-based delay sensor

As seen in Figure 3.12, the digital sensor previously proposed requires lots of logic resources and therefore area to measure performance of a certain chip region. This solution might not be feasible for a system with multiple ABB or AVS domains since power consumption and area penalty would be prohibitive. This fact motivates the design of a light-weight delay sensor which converts delay shifts into analog voltages.

In this type of sensor, the signal coming from an external pulse generator passes through an N-stage VCDL that delay the reference pattern pulse according to PVTA variations. The reference signal pulse width corresponds to the nominal end-to-end delay of the VCDL in the absence of any variation for a given operating frequency. The reference pulse frequency is set based on the desired response time of the sensor such that a faster response time implies a higher reference signal frequency at the expense of power consumption overhead.

Once the reference signal passes through the VCDL, both, delayed and reference signal arrive at the phase detector. This phase detector achieves its steady state when the VCDL output rising edge is produced at the same time as the reference signal falling edge, and thus VCDL nominal end-to-end delay (corresponding to the reference pulse width) is achieved. The phase detector generates two output signals: the flag that determines if the VCDL is faster or slower than the nominal case, and a pulse where the timing difference between reference and current delays is encoded.

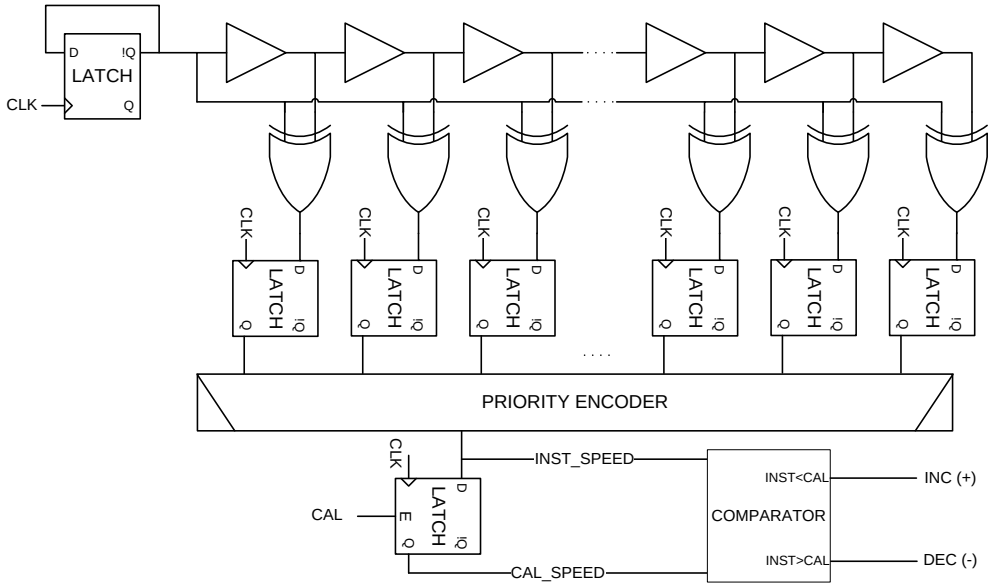

Figure 3.12 Schematic of the proposed TDC-based delay sensor.

Figure 3.15 shows a chronogram where an example of the three possible states is represented. It can be seen that *ERROR* flag rises when VCDL end-to-end delay is smaller than the reference pulse width that, in turn, is equivalent to the system clock cycle. To indicate that the circuit is faster than the nominal case, *SIGN* flag is set to low. Notice that *ERROR* flag encodes the size of the timing difference in its pulse width, so that the higher the difference, the wider the pulse. Similarly, when VCDL is slower than expected *ERROR* flag rises but *SIGN* flag is set to high. When VCDL has an end-to-end delay very similar to the expected one, *ERROR* flag will no longer rise.

3.6 Concluding remarks

According to simulations, transistor channel length variations are the main source of leakage variability while power supply voltage variations are the main cause of dynamic power variability. On the other hand, transistor delay variability depends on both channel length and voltage variations. This may explain the fact that delay variability maintains a good correlation with both static and dynamic power consumption variability. This good correlation with the other magnitudes makes delay a good candidate as an observable of PVTA variability.

As for the feasibility of the proposed sensors, thermal measurement of dynamic power consumption could be implemented in a very reduced area with an acceptable dynamic range. Moreover, this dynamic range could be improved by replacing standard cell inverters of the ring oscillator by denser full custom inverters. Conversely, leakage power thermal sensing is not feasible since it exhibits a poor dynamic range and a large area penalty. In case leakage sensing was considered as the observable magnitude to monitor PVTA variations, an alternative sensor design should be proposed [13][77].

Two delay sensor proposals were presented. Despite both proposals are valid and work as expected, in practice the TDC-based delay sensor proposed has a large area overhead and power consumption, while the VCDL-based solution provides similar delay variability information with much less hardware requirements.

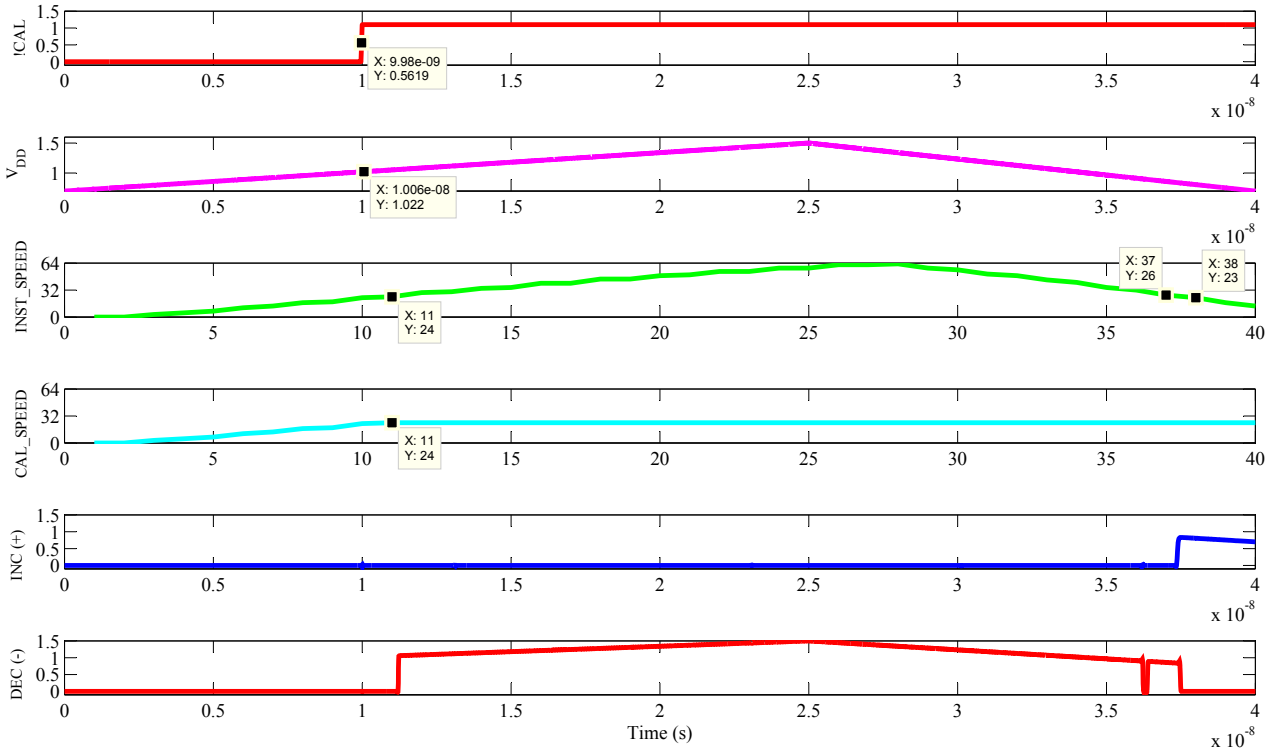

Figure 3.13 TDC-based delay sensor chronogram.

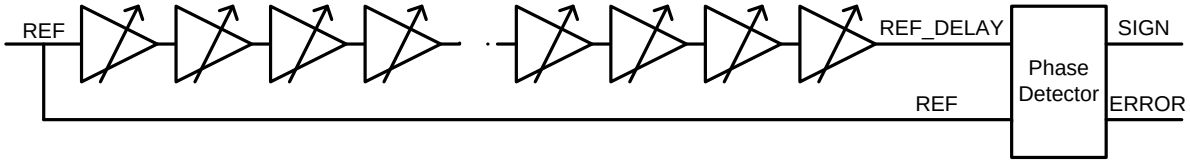

Figure 3.14 Schematic of the proposed VCDL-based delay sensor

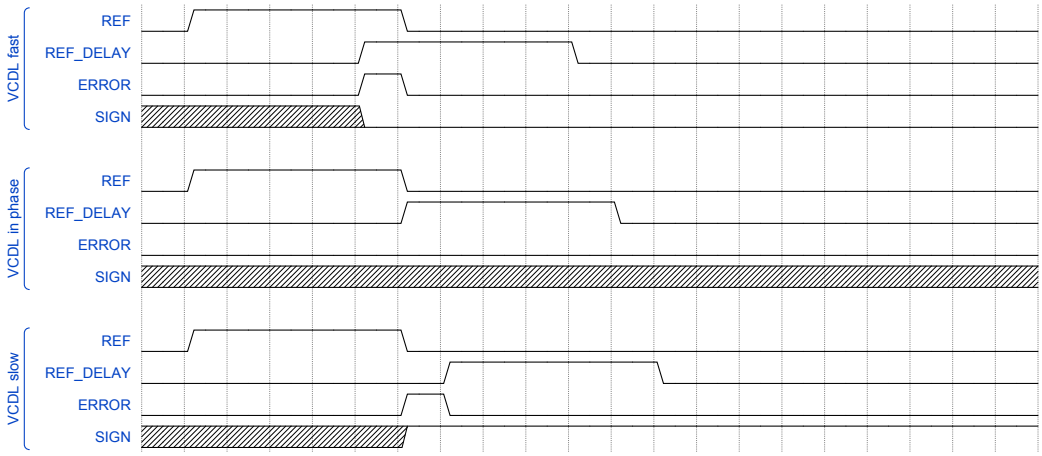

Figure 3.15 VCDL-based delay sensor chronogram.

CHAPTER 4

POST-SILICON TECHNIQUES STUDY

The second remarkable contribution is a study of the effectiveness of ABB and AVS when both are applied simultaneously. The objective pursued is the achievement of maximum tuning range and thus to ensure that even those circuits more penalized by variability are able to recover their nominal behaviour and thus maximize Parametric Yield.

In order to check the effectiveness of this method PVT variations have been simulated by combining the Monte Carlo method with an iterative algorithm that finds the optimal combination of substrate bias and power supply voltages to optimize either static power, dynamic power or performance. Moreover, side effects produced by the optimization in the non-optimized magnitudes are also studied.

As seen in Chapter 2 there are many reactive ways to compensate the effects of variability and they can be classified as follows:

- Modification of electrical parameters of the circuit by tuning either substrate voltage (ABB) [13][14][15][16] or power supply voltage (AVS) [71][17][20].
- Clock frequency scaling [18].
- Implementation of circuits tolerant to timing violations without losing data [73][74].

This thesis is focused on those techniques based on the electrical tuning of circuits, i.e. ABB and AVS. Similarly to [17], in the previous chapter sensitivity of power and delay to ABB and AVS is studied. The main difference with the cited study lays in the fact that 2D level curves are extracted and thus it is easier to understand how to combine both techniques to minimize variability.

Then, once ABB and AVS tuning mechanisms are well understood, circuits under the effects of PVT variations are simulated using Monte Carlo analysis, and they are optimized using such techniques and assuming different sensing approaches. An iterative algorithm repeats each Monte Carlo sample simulation until the optimal ABB and AVS are found. Finally, the obtained results with all the sensing approaches are compared.

4.1 Impact of ABB and AVS on variability indicators

The aim of this study is to understand how power consumption and delay evolves with ABB and AVS techniques. The full theoretical analysis can be found in Appendix A [56].

To obtain the relationship between ABB, AVS and variability indicators, 2D parametric simulations are performed using a chain of 15 inverters as a testbench. The models used for this test are the BSIM4 device models for 65 nm technology with the typical corner configuration.

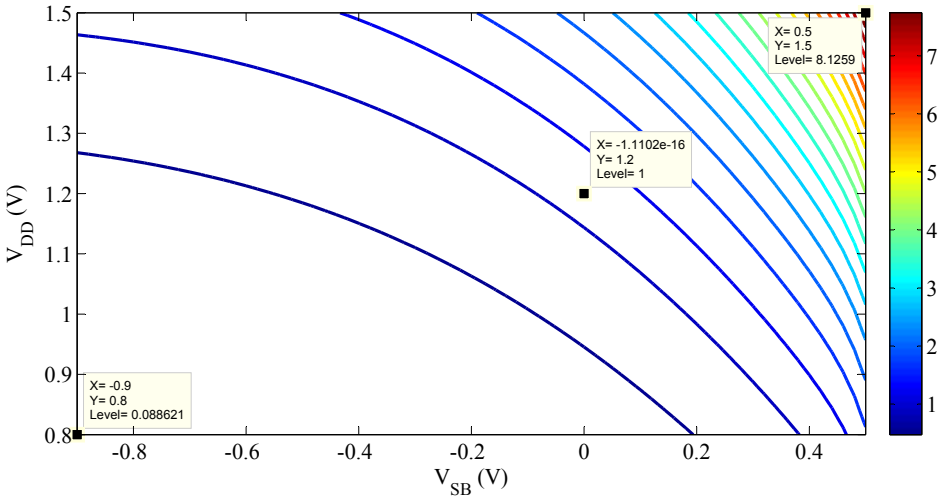

Figure 4.1 Normalized static power consumption as a function of ABB and AVS.

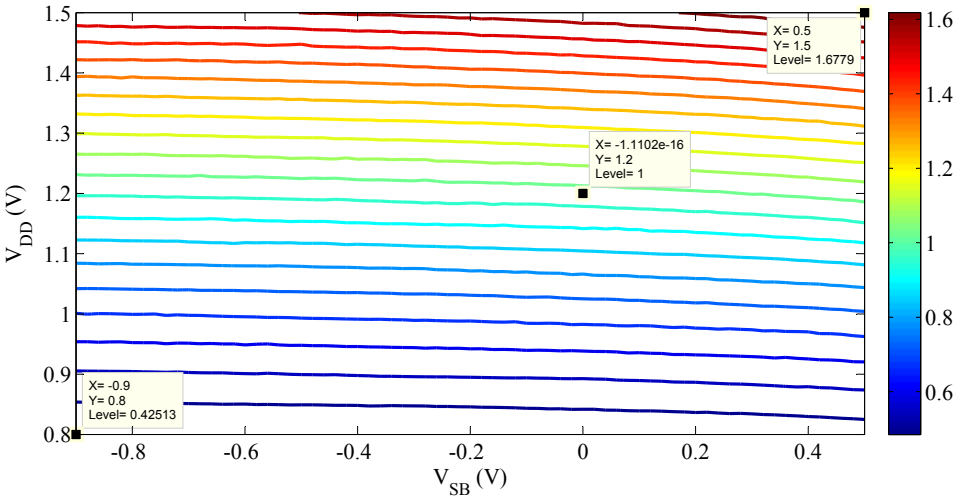

Figure 4.2 Normalized dynamic power consumption as a function of ABB and AVS.

From the level curves in Figures 4.1 to 4.3, both static power and delay increase can be observed either with V_{DD} or V_{SB}. On the other hand, dynamic power almost only depends on V_{DD}.

Static power consumption is highly sensitive to ABB and AVS since applying RBB and the minimum supply voltage, leakage is reduced by a factor greater than 10, while the maximum FBB and supply voltage produce a leakage 8 times larger than the nominal case. This means that most of the Monte Carlo samples with large leakage variability can be corrected.

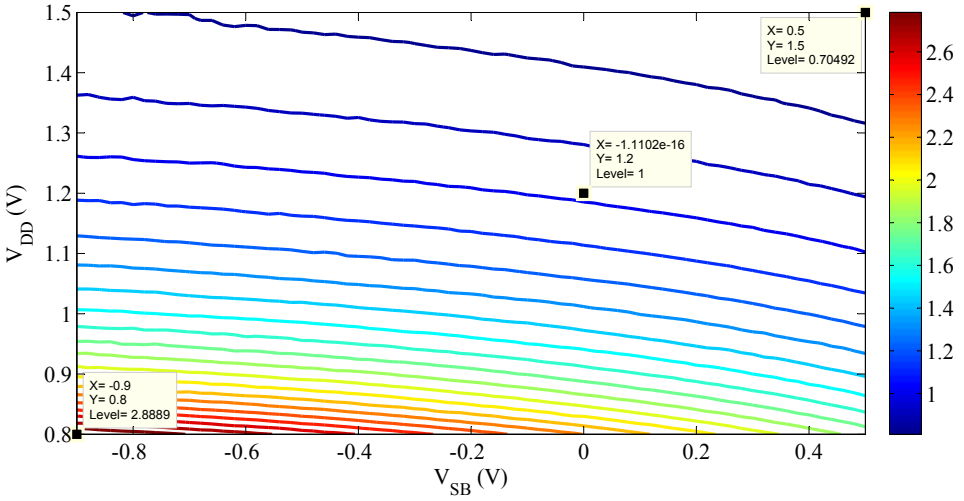

Figure 4.3 Normalized transistor delay as a function of ABB and AVS.

Dynamic power consumption mostly depends on power supply voltage. Hence, assuming that dynamic power is used as the observable magnitude, it would not make sense combining AVS and ABB since the sensor would not be able to distinguish whether ABB is effective or not.

Concerning delay, circuit latency can be reduced up to a 30% by applying the maximum V_{SB} and V_{DD} (0.5 V and 1.5 V respectively), while applying the minimum V_{SB} and V_{DD} (-0.9 V and 0.8 V respectively) increases latency by a factor 2.9.

4.2 Optimizing circuits by means of ABB and AVS

Once seen the statistical behavior of power and delay in the presence of PVT variability and having introduced ABB and AVS post-silicon tuning techniques, the present Chapter studies the impact of ABB and AVS in the optimized magnitude and also in the non-optimized magnitudes. The aim of this study is to understand the impact of the optimization of one observable magnitude on the overall variability.

Simulations were done as follows (see flow chart in Figure 4.4): the i-th Monte Carlo simulation sample is performed using nominal V_{DD} and Zero Body Bias (ZBB). From the first iteration of the i-th sample simulation results are computed. These results are static power, dynamic power and delay. Then, from this obtained results new tuning parameters for V_{DD} and V_{SB} are computed. New tuning parameters will be those that reduce variability of a given magnitude (bringing the magnitude closer to the nominal values). I-th Monte Carlo simulation is then performed again with the new electrical parameters and circuit is measured again. This procedure ends when the previous and the new computed V_{DD} or V_{SB} values are equal within a given resolution ($\pm 0.01 V for \mathrm{V}_{SB}$ and $\pm 0.005 V for \mathrm{V}_{DD}$) or minimum / maximum tuning values are set.

To perform the current analysis variability sensors are considered to be ideal. Simulations are performed considering the following sensing scenarios:

1. Static power sensor.
2. Dynamic power sensor.
3. Delay sensor.
4. Dual sensor: delay sensor plus power sensor.

For each of the sensing scenarios 250 Monte Carlo samples of an 8-Bit RCA are simulated applying $\pm 3\sigma$ parameter variability shown in Table 3.1, and using the algorithm described in Figure 4.4. Figures 4.5 to 4.7 show

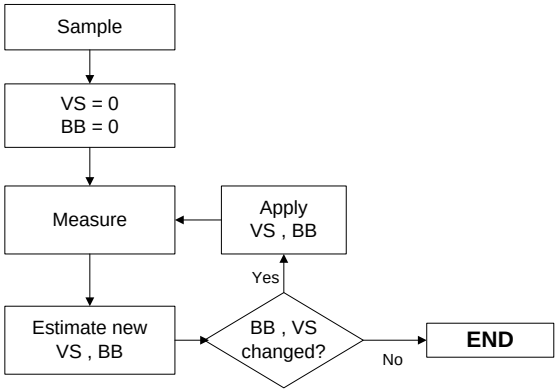

Figure 4.4 Variability optimization flow chart.

Parameter	Nominal	Min	Max
V_{SB}	0 V	-0.9 V	0.5 V
V_{DD}	1.2 V	0.8 V	1.5 V

Table 4.1 Ranges of ABB and AVS voltages.

simulation results. The blue bold line corresponds to the scenario where no variability corrections are applied. Notice that this trace is exactly the same as the ones shown in Figures 3.4 to 3.6.

- **Static power optimization:** although this variability reduction strategy achieves excellent results in terms of leakage power, this reduction of leakage variability implies a worsening of statistical performance in terms of delay and dynamic power consumption. Therefore, this sensing strategy implementation would worsen Parametric Yield loss rather than improving it. This can be attributed to the low correlation between leakage power and delay and, in particular, between leakage and dynamic power.
- **Dynamic power optimization:** this method does not contribute to reduce PVT effects substantially. Despite dynamic power variability is reduced, delay and static power do not have remarkable improvements.
- **Transistor delay optimization:** this method reduces delay variability without worsening dynamic power consumption variability. Leakage power variability is also substantially improved.
- **Dual optimization:** dual sensor strategy achieves the best trade-off between dynamic power and delay. Indeed, in 90% of the simulated samples dynamic power consumption and delay is below +5% nominal values. However, leakage power consumption experiences an increase in terms of mean and variance.

In summary, each type of sensor can be used to decrease variability of the magnitude being observed, but typically worsening the non-observed magnitudes. Static power sensing is a clear example of this phenomenon. None of the sensing strategies were able to reduce the variability of all the indicators, but the delay sensor is able to improve two of them (leakage and delay) without significantly worsening the third one (dynamic power). Finally, dual sensing strategy significantly reduces the effect of variability in dynamic power and delay at the expenses of leakage power consumption. However, leakage current variability is not necessarily an issue. Static power consumption only becomes an issue when stringent power restrictions exist and, what is more, contribution of static power to the overall power consumption is only critical when circuits are idle. Thus, some authors in literature [78] propose mechanisms to reduce only leakage during idle periods.

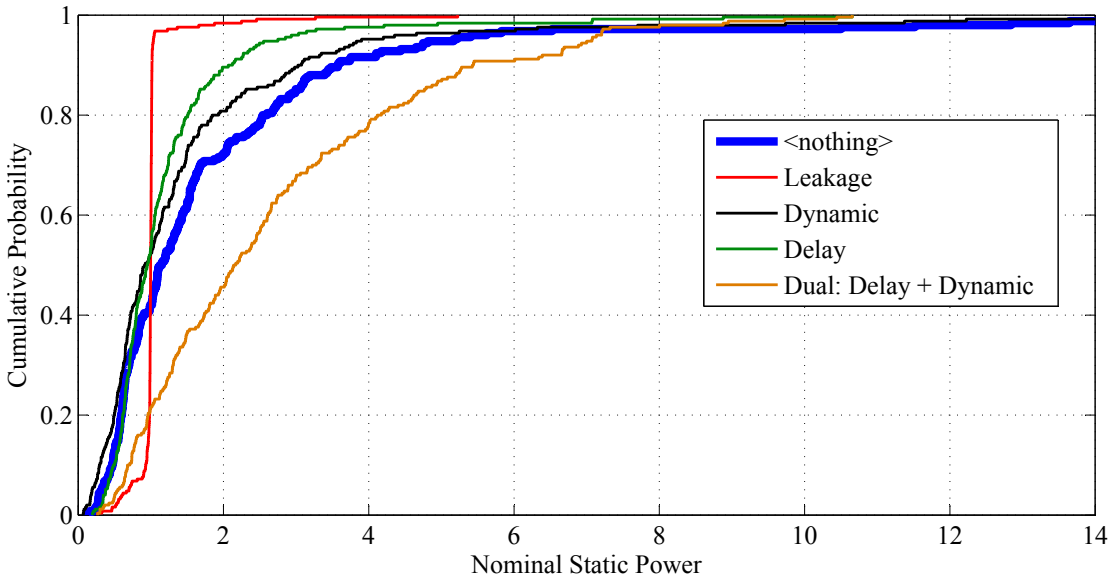

Figure 4.5 Normalized static power CDF in an 8-Bit RCA applying ABB and AVS.

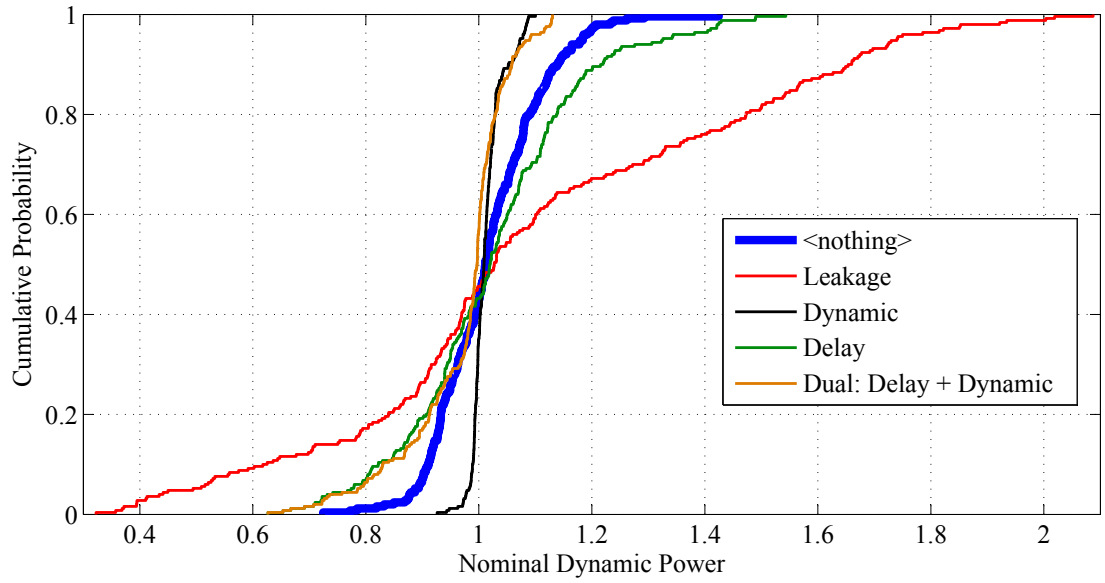

Figure 4.6 Normalized dynamic power CDF in an 8-Bit RCA applying ABB and AVS.

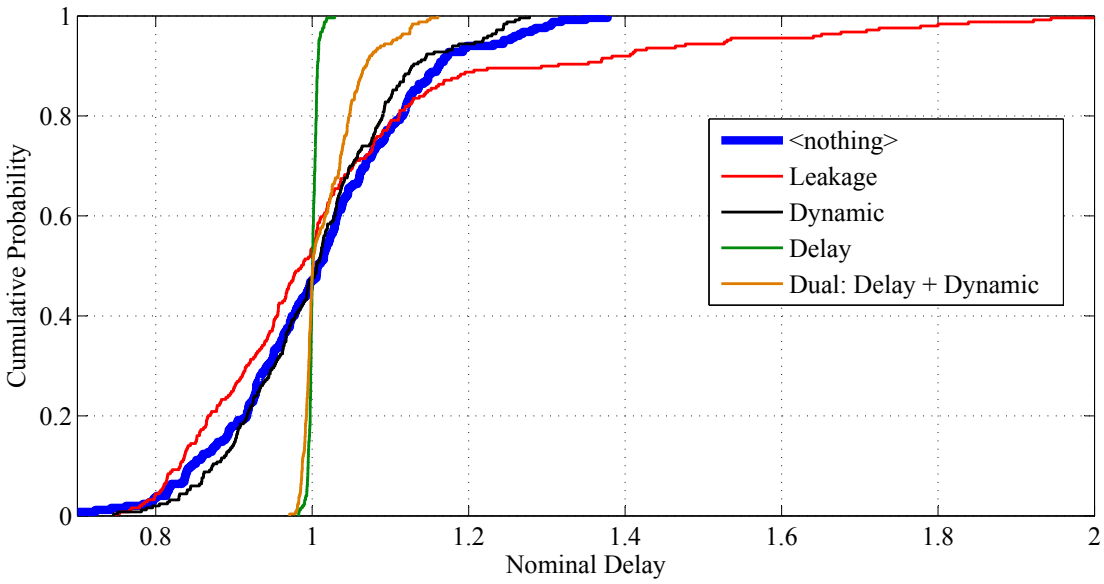

Figure 4.7 Normalized transistor delay CDF in an 8-Bit RCA applying ABB and AVS.

CHAPTER 5

VARIABILITY MEASUREMENTS IN A 40 NM IC PROTOTYPE

The results obtained in previous chapters are backed by circuit simulations of device models provided by manufacturers. These models are very useful to predict the behavior of circuits before being physically implemented. Moreover, variability effects can be simulated so that Parametric Yield can be estimated in advance.

However, some spatial scale variability effects such as WID and D2D variations are very hard to predict since the simulator does not take into account proximity effects arising from physical implementation of circuits. The need of input data for modeling and for verifying methodologies and concepts motivated the physical implementation of a demonstrator under the framework of MODERN ENIAC project. This demonstrator IC prototype was done using ST 40 nm technology.

On the other hand, this demonstrator was also used as a proof of concept for other research lines in the HIPICS research group. More concisely, to the research line in regular layouts. The proposed demonstrator has also the objective of showing the improvement of regular layout variability, as a consequence of a better result of lithographic translation of the layout patterns and a corresponding reduction in channel length variability. Despite that the study of layout techniques is out of the scope of this thesis, the combination of both layout techniques and post-silicon techniques is considered as a way to relax the adaptive techniques tuning range.

The Via-Configurable Transistor Array (VCTA) proposal, described in [37], is a very regular structure that maximizes layout regularity by setting up regular interconnects and enforcing all transistors to have the same dimensions. It is based on a basic cell that contains 6 PMOS and 6 NMOS transistors, and an interconnection grid from M1 to M3 levels (see Figure 5.1). The connectivity of the transistors is done by proper use of contacts and vias.

5.1 Demonstrator chip description

The circuit under test for this IC demonstrator is a 64-stage VCDL. A VCDL is a digital delay line the speed of which can be modified by acting on the input control voltage V_{CNT} of the circuit. The advantages that bring this circuit are many: it is a compact, repetitive and modular design, which considerably reduces design time. In

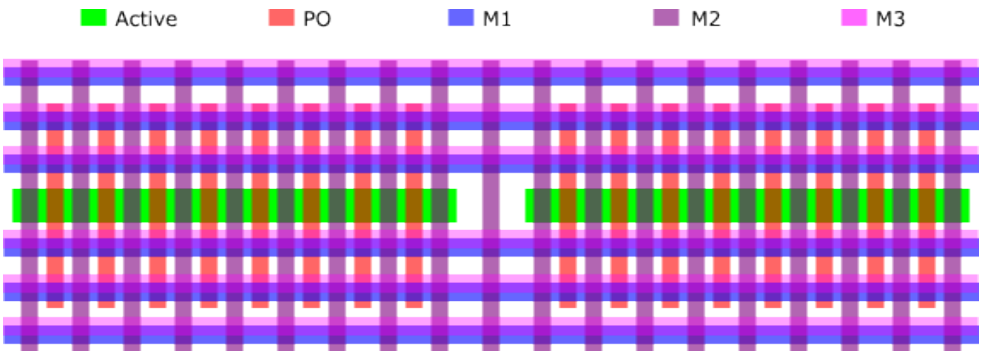

Figure 5.1 Basic VCTA cell.

addition, the VCDL circuit eases mismatch measurements since WID and D2D variability effects can be uncoupled by acting on V_{CNT} and thus fixing VCDL end-to-end delay to a certain value.

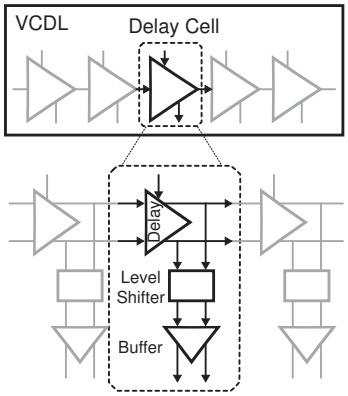

Figure 5.2 Schematic of the VCDL.

Figure 5.3 shows VCDL block diagram and multiplexing scheme of a single design style. The outputs of each VCDL are multiplexed to reduce the number of IC pin outputs. More concisely, any of the 64 stages that form a VCDL can be connected to the output through the origin or destination multiplexer. Finally, origin and destination signals from each VCDL are multiplexed together and thus they provide the output of a single origin and destination signals per design style. In order to increase the statistics and also to be able to measure the effect of WID variations, 8 VCDLs are placed one next to each other.

In order to evaluate the differences between regular and non-regular layouts, two layout implementations were created. We designed a VCTA implementation of a VCDL and we also designed a non-regular full custom (FC) design, more sensitive to lithography effects. The transistor channel area is the same for VCTA and FC styles. Furthermore, the VCDL was implemented in two sizes for each design style, since according to Pelgrom's law random variability decreases with $\sqrt{WL}$. For each design style, VCDLs are implemented in two versions: one with transistors having the minimum channel width (FC x1, VCTA x1) and another one with transistor width of four times this minimum channel width (FC x4, VCTA x4). In this way, it will be possible to determine the dependence of local variations as a function of transistor size and thus determine the relative influence of layout dependent variations against random variations.

As seen in Figure 5.4, every chip contains 8 VCDL instances of each layout style and size. As the overall chip area was constrained to 1 mm^2, a trade-off existed between the number of VCDLs and the number of stages per VCDL. As reported in [79], longer delay lines suffer larger local variations (mismatch), at the expense of the statistical population. The number of available silicon dies was 22, and hence, the total VCDL population is 176 for each layout style and transistor size. Further details can be found in Appendix B [57].

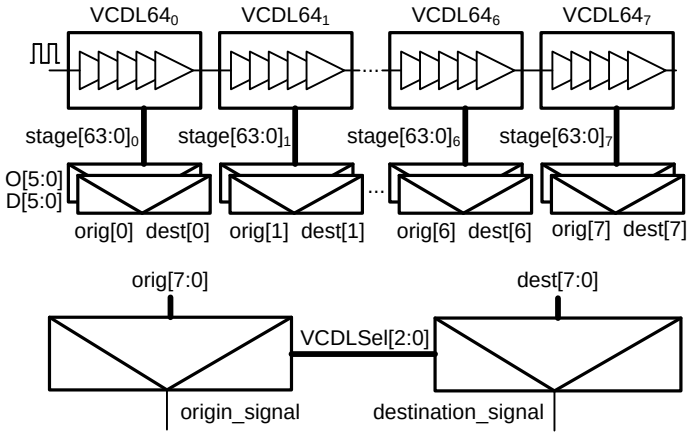

Figure 5.3 VCDL multiplexing scheme. *O[5:0]*, *D[5:0]* and *VCDLSel[3:0]* multiplexer selection signals are externally controlled.

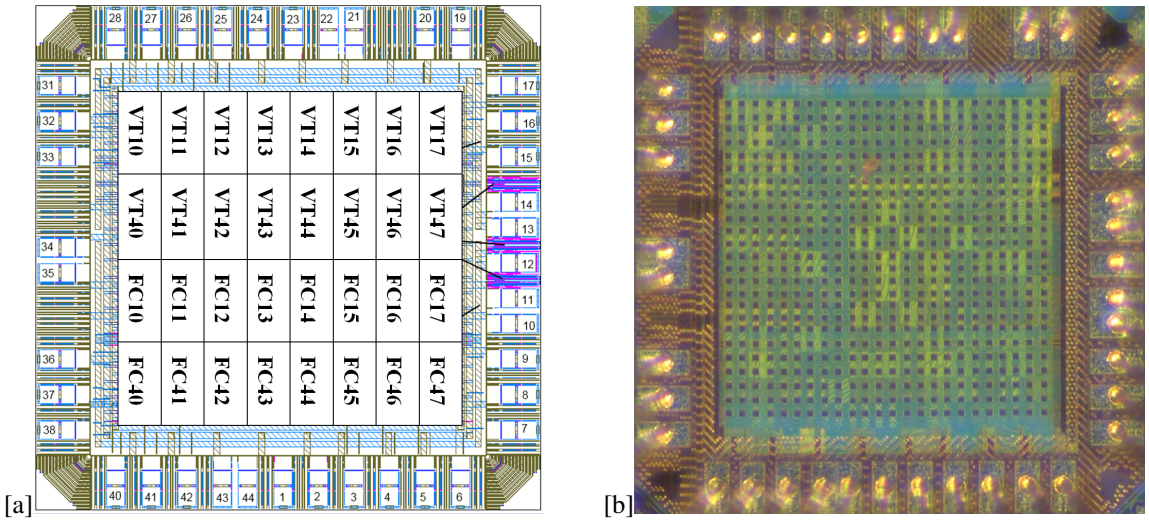

Figure 5.4 [a]: chip floorplanning. In the labels the first two characters indicate whether the block is full custom or VCTA; the next one indicates the layout size and the last one indicates the instance number. [b] chip picture.

The chip measurement setup is shown in Figure 5.5. A pulse generator provides a square pulse which is generated as the reference input signal for the chip. Chip V_{DD} and VCDL V_{CNT} are generated by means of a controllable DC power supply. Notice that V_{DD} is a fixed voltage while V_{CNT} can be adjusted to compensate WID and D2D variations. Multiplexer signals are controlled by a PC which, in turn, acquires timing measurements from a high-end oscilloscope that monitors VCDL multiplexed outputs. Chips were tested in the test board shown in Figure 5.5. Global and local variations measurement procedures are described in detail in Appendix B [57].

5.2 WID and D2D variability measurements

Global effects of variability were obtained by measuring end-to-end delay of the entire VCDL population. Figure 5.6 shows D2D histograms for all the layout styles and types, where the same bin color is used for the VCDLs corresponding to a given instance number. This data representation shows a clear correlation between VCDL

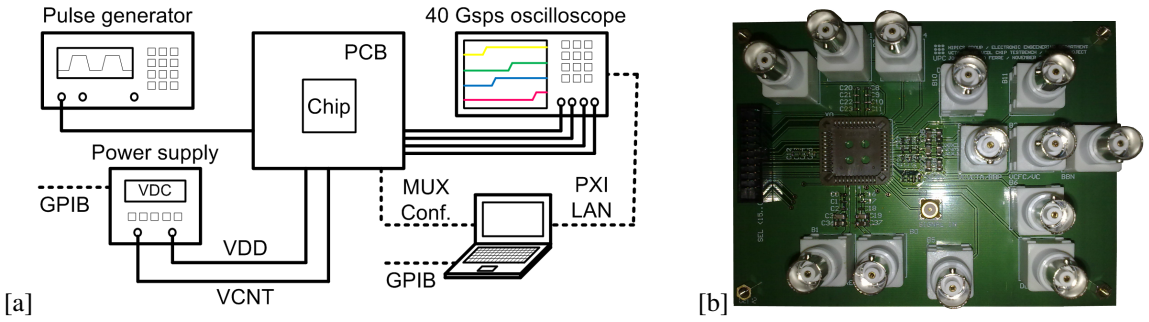

Figure 5.5 [a]: testbench setup diagram. [b]: demonstrator chip test board.

physical placement and end-to-end delay, and thus indicating the presence of WID variability. From the obtained results the following conclusions are drawn:

- **Overall variability:** minimum size FC VCDLs are the most sensitive circuits to variability ($\sim$23%). Notice that the obtained results are quite compatible with variability simulation results showed in Chapter 3, taking into account that chip measurements only consider process and voltage variations (measurements were done in stable temperature conditions). It can be seen that overall variability is improved ($\sim$15%) in FC designs by increasing transistor width. On the other hand, VCTA layout designs significantly reduces variability ($\sim$9%) and no significant improvements are obtained by increasing the size of transistors. The conclusion is that most of the variability is produced by systematic effects.

- **D2D variability:** it does not significantly improve with the increase in transistor width (see Figure 5.6 legends). Indeed, D2D variability slightly increases in FC x4 layouts. It can be concluded that regular layouts are qualitatively less sensitive to D2D variations than non-regular ones.

- **WID variability:** it is clearly the dominant effect of variability in non-regular layouts. Timing differences between VCDL samples belonging to the same chip are higher than the same VCDL samples belonging to different chips. This leads to differences between VCDL within the same chip of 20% for the case of FC x1 samples. Figure 5.7 better reflects the importance of WID variations. This picture shows the average of the normalized VCDL end-to-end delay for each of the implemented instances of the chip. VCDL instances closer to the die boundary experience smaller delays than the ones in the middle. This phenomenon can be partly attributed to a non-uniform power supply distribution as the effect can also be seen in VCTA layouts, but it is clearly aggravated by the use of non-regular layouts.

As commented at the beginning of the chapter, the combination of layout techniques such as VCTA with adaptive techniques is also an object of study in this thesis. One of the advantages of this combination is that the required tuning range of the adaptive circuit would decrease since variability is partly mitigated by design.

To compare tuning range requirements between VCTA and FC layout styles, all the VCDLs were adjusted to have 10 ns of end-to-end delay. As the demonstrator chip did not include any adaptive circuit implemented on it, the adjustment of V_{CNT} was externally done by the test system. Figure 5.8 shows the histograms with the configured V_{CNT}. Notice that there are huge differences between the average V_{CNT} configured in each layout style and size. VCTA x1 and VCTA x4 require a V_{CNT} 17% and 37% higher than FC x1 respectively. This can be explained by the fact that VCTA have performance losses associated to the large parasitic coupling capacitance due to dense and long parallel interconnection lines. In contrast with FC designs, the wider the transistor width, the higher the delay since parasitics are the dominant effect.

Nevertheless, the interesting parameter for this study is the required range of V_{CNT} in percentage, which decreases by half.

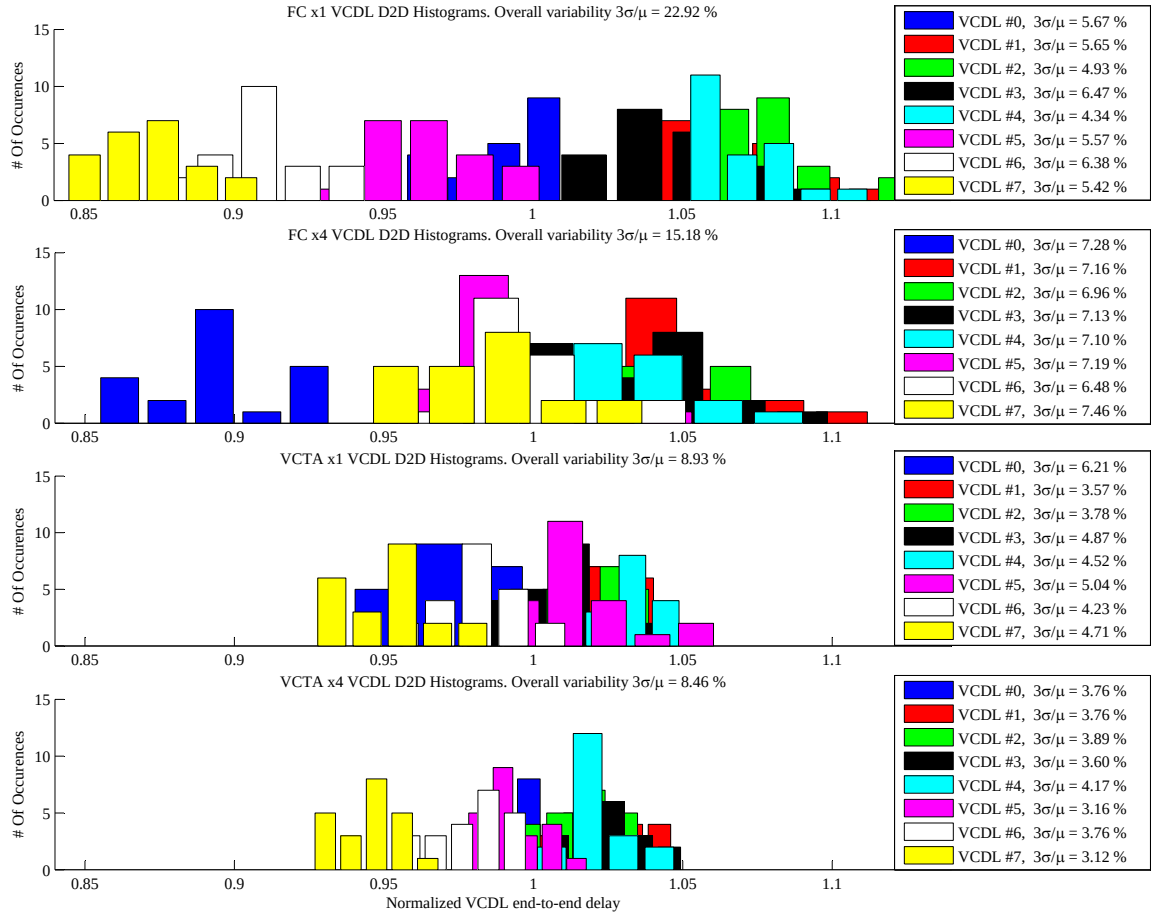

Figure 5.6 D2D histograms of the entire VCDL64 population for all the layout styles and sizes.

5.3 Transistor mismatch measurements

As seen in Chapter 1 the problems of transistor mismatch become prominent as device sizes are driven further into sub-micron geometries. In contrast with D2D and global WID variations, the main feature of mismatch (local WID variations) is the unpredictable variability pattern between contiguous devices. Such kind of variability is caused by atomistic effects of the device structure and dimensions (random), and on others related to manufacturing process imperfections (systematic). While the first is inherent to transistor scaling, the second one can be improved by means of layout techniques. The goal of the following study is to measure random and systematic effects of mismatch to determine the effectiveness of regular layouts in the mitigation of systematic mismatch.

In order to decouple global effects of variability and also to compare mismatch measurements between all the VCDL designs, VCDL end-to-end delay is adjusted to 10 ns by tuning V_{CNT}. In this way, timing uncertainties between the expected cumulated delay and measured delay will decrease as the measured stage gets close to stage 0 and stage 63. Mismatch variability is measured by computing the Integral Non Linearity (INL) of each VCDL 5.1, i.e. the difference between the measured cumulated delay and the expected cumulated delay of each stage. Then, standard deviation is computed per VCDL stage so that those stages with higher timing differences from VCDL to VCDL will also exhibit higher INL.

$$\boldsymbol{INL(j,k) = Delay(j,k) - k \cdot \frac{10ns}{63}} \tag{5.1}$$

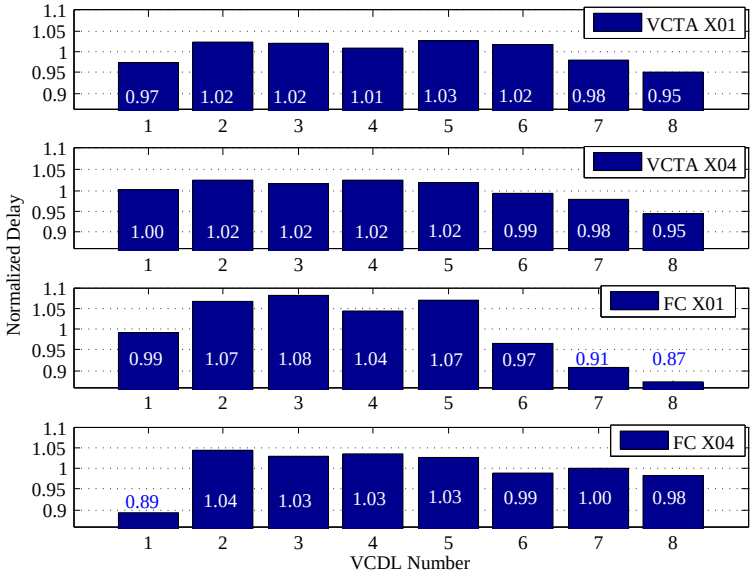

Figure 5.7 Normalized average path delay for each VCDL instance in the chip.

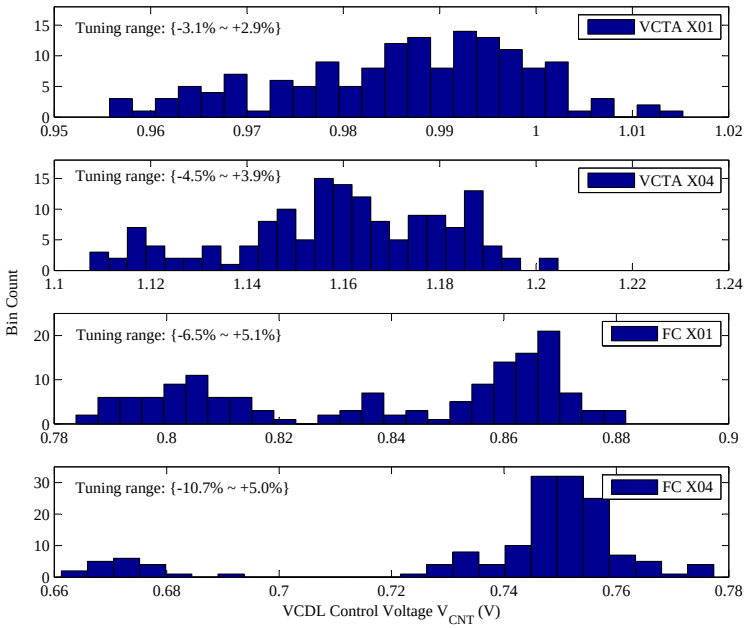

Figure 5.8 Histogram of the configured V_{CNT} so that end-to-end delay is 10 ns for each VCDL instance in the chip.

$$\sigma_{INL}(k) = \sqrt{Var[INL(j,k)]} \tag{5.2}$$

It was already known from simulations that the maximum expected variability should appear in the middle VCDL stages. However, the non-uniformity of process variations along the VCDL as well as multiplexer variability, supply voltage noise and oscilloscope resolution produced distortions in the measured σ_{INL}. To compensate these effects, measurements were fit by means of a half-ellipse curve. Further details can be found in Appendix B [57].

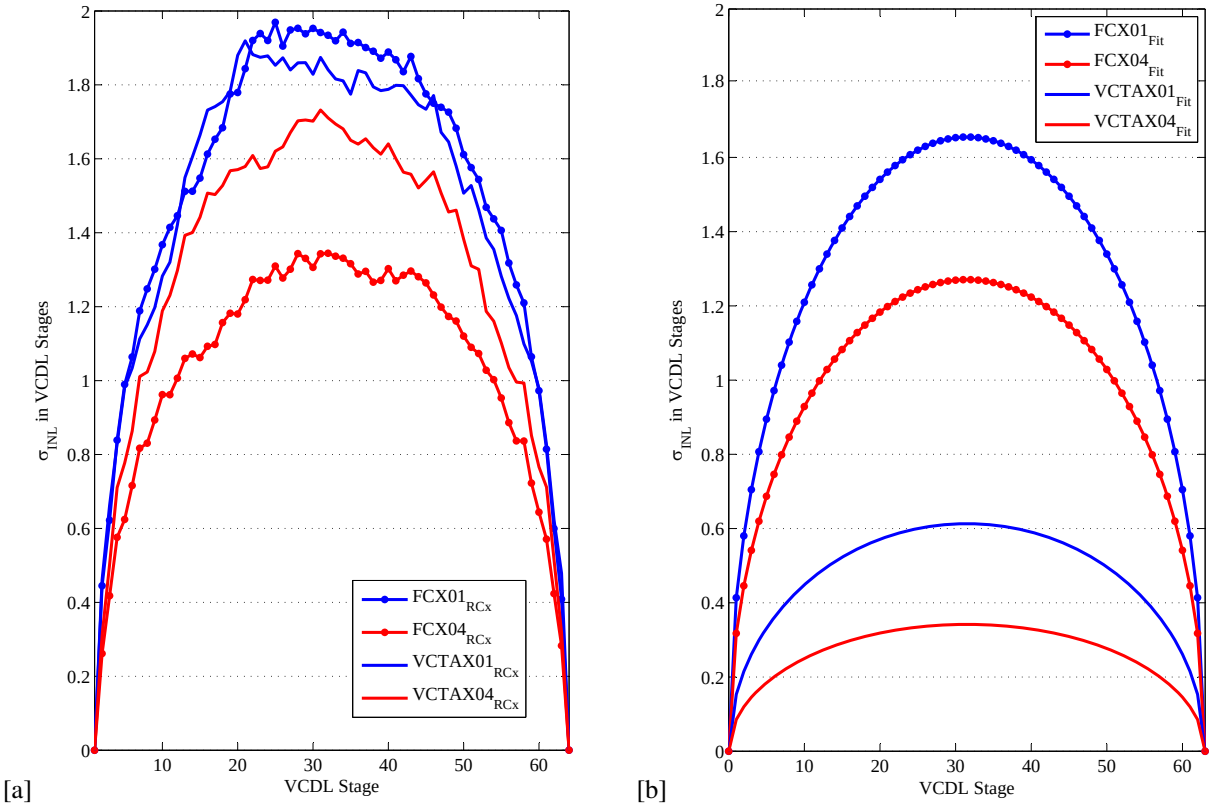

Figure 5.9 Estimation of σ_{INL} due to local variations (mismatch). [a]: RC parasitic extracted layout simulations, [b]: curve fitting of the demonstrator chip measurements.

Figure 5.9 shows mismatch variability results obtained from RC parasitic extracted layout simulations (left) and half-ellipse curve fitting of the chip measurements (right). It can be seen that FC design styles agree with simulations while variability in VCTA design styles is clearly overestimated by simulations. This can be explained by the fact that device model always considers mismatch as a mixture of random and systematic effects, and therefore layout manufacturability optimizations are not taken into account. This fact explains why VCTA x1 and FC x1 exhibit similar mismatch variations from simulation results.

Observing measured variability in VCTA layouts we can conclude that random variability is the main cause of local variations in regular layouts, since increasing 4 times the layout size, variability improves by a factor of 1.8, close to factor 2 which is the expected improvement in random variability according to Pelgrom's Law. On the other hand, in full custom layouts local variations are dominated by systematic effects, and consequently, variability only improves 30% when transistor width is increased by a factor of 4.

In conclusion, regular layout designs suffer less from variability than non-regular designs, and hence reliability is improved with this design methodology. The price to pay is the area penalty: VCTA implementations of the VCDL have an area overhead of 60% in comparison with their full custom counterparts. Note that this overhead depends on the use of transistors, and this is circuit dependent. In this particular case, the transistor utilization is 70%. However, there are other regular layout proposals [9] with less area penalty at the expenses of regularity.

It is clearly demonstrated that regular layouts relax the required tuning range of adaptive techniques. Consequently, it is feasible to implement a mixed solution combining a low-area adaptive post-silicon tuning system (either by means of ABB or AVS) and layout techniques.

CHAPTER 6

BODY BIAS GENERATOR CIRCUIT PROPOSALS

To conclude this thesis, the last contribution has been the design of a Body Bias Generator which automatically adjusts substrate voltage according to local on-chip variations, and thus reducing the effects of PVTA variability. The main virtues of this proposal in comparison with other existing works in the literature are compactness, high tuning range and low power consumption. These characteristics together with the fact that this BBG works in closed loop and hence it does not require external control, make feasible the integration of hundreds or even thousands of BBIs in a single IC. Each island has its own BBG which is adjusted according to the local PVTA variations of that chip region.

Once seen the effects of variability in simulations as well as in chip measurements, and having studied post-silicon tuning techniques, a couple of circuits to reduce the effects of variability are proposed. The criteria used to make the choice of the observable variability indicator and tuning mechanism are the following:

- **Feasibility of implementation:** the proposed circuits should not only be a mere academic solution. Therefore, they have to be simpler enough as to be integrated into EDA tools.
- **Low area and power penalty:** Parametric Yield improvement has to be clearly higher than the area overhead produced by inserting our circuit into the chip. Otherwise, the effort done in the integration of proposed solutions is not worthwhile. Similarly, the increase in power consumption produced by the circuit should be as small as possible. It is difficult to establish an arbitrary measure to determine whether power consumption overhead is affordable or not. As a first approximation we can consider that power density must not increase by the insertion of the circuit, that is to say that power consumption of the circuit must be less or equal to the product of area overhead and power density.
- **Scalability:** proposed solution should be valid either for small (less than 1 mm^2) or big (less than 1 cm^2) designs. Moreover, proposed circuits should also scale with transistor scaling.

We chose delay as the observable variability indicator for the following reasons:

Body Bias Voltage	Static Power	Dynamic Power	Max Frequency
-1.8 V	-87.7%	-15.5%	-29.7%
-0.9 V	-71.1%	-7.7%	-15.2%
0.35 V	+87.7%	+3.1%	+6.2%
0.7 V	+196%	+6.3%	+12.6%

Table 6.1 Body Bias adjust range of a 32-stage VCDL in 28 nm FD-SOI technology. Maximum RBB (-1.8 V) substantially reduces static power at the expenses of transistor delay. On the contrary, maximum FBB (0.7 V) improves transistor performance and dramatically increases static power.

- Timing failures are the most critical source of reliability loss. While chips exhibiting thermal hot spots or suffering high leakage currents can still be functional, chips containing circuits that suffer from clock violations cannot.

- As seen in Chapter 3, transistor delay variability maintains a certainly good correlation with both static and dynamic power. This is then confirmed when overall variability is optimized using delay as observable magnitude (see Chapter 4).

- From the studied sensor implementations, the VCDL-based delay sensor is the one that requires less hardware and power consumption.

- Dual sensing implementation is discarded. Despite it can potentially reduce both dynamic power and transistor delay variability, the complexity associated with combining information from two sensors and the increase in area overhead, make this approach difficult to implement.

Notice that the main drawback of using delay as observable magnitude is the insensitiveness to thermal hot spots that might jeopardize ICs.

As for the post-silicon tuning mechanism we chose ABB instead of AVS or a combination of them for the following reasons:

- As seen in Chapter 5, and more precisely in Figure 5.7, the effects of WID variations in contiguous VCDLs may differ substantially. Therefore, applying a single body bias or supply voltage to the chip does not avoid local differences. Instead, the die may be partitioned into a certain number of islands with independent control.

- The use of AVS would limit the number of islands since each island would require a tunable voltage regulator with driving capability enough to supply the transistors of that island.

- ABB has not that limitation since, in principle, bulk terminals are of high impedance.

- The effectiveness of body bias is known to decrease with nanometer bulk MOS devices. However, with the appearance of FD-SOI technology with ultra-thin body and buried oxide (UTBB), body bias is again seen as one of the best ways to control variability [80]. FD-SOI allows a wide range of body bias voltages and consequently a larger control on the transistor threshold voltages (see Table 6).

The objective is thus to minimize the effects of PVTA variability by measuring delay and applying the body bias voltage that brings the circuit to its nominal performance. To compensate local effects of variability we consider the BBI approach proposed in [66][36]. This concept consists in partitioning silicon dies into islands to control different regions of the chip with different bias voltages.

In the approach considered in this proposal (see Figure 6.1), each BBI has a sensor and a BBG with a self-adaptive procedure to adjust the body bias voltage at runtime so that the circuits contained in the BBI operate near to nominal conditions. The adaptation is local to each BBI and decentralized, so the overhead is only related to the sensor and BBG contained in the BBI.

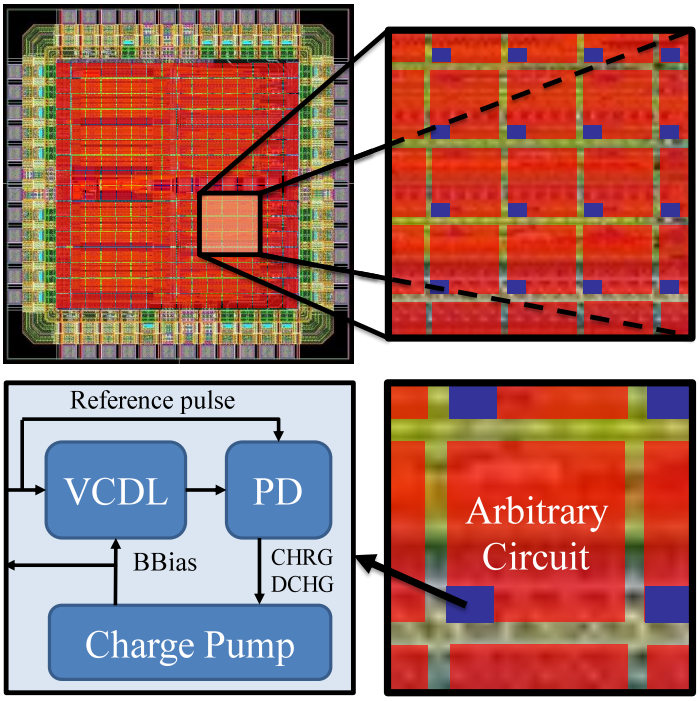

Figure 6.1 Chip is partitioned into small body bias islands. Each island has a compensation circuit comprising a delay line, a phase detector and a charge pump.

As previously commented, two BBGs were proposed. The first one is a simple and light-weight BBG with FBB capability only. The second one has a more complex design which allows both FBB and RBB modes. The main features of these two proposals are the following:

- Circuits are almost fully analog to reduce area penalty, power consumption and response time. This allows us to dispense with DACs.
- Sleep mode to minimize leakage when island is idle.
- BBG works in closed loop based on a DLL-like structure so that the measured delay drift in each BBI is used to track variations by modifying body bias output.
- Scalable and easily integrable into an automated design flow since the architecture is fully distributed.

6.1 FBB Generator circuit proposal

This was the first BBG circuit version which was able to generate voltages only within power supply voltage rails. The main advantage of such implementation is the low area penalty (81 μm^2). Of course, the main disadvantage is the low adjust range (700 mV) that limits the range of compensable delay variations. However, this low tuning range could be enough for designs where systematic process variability has been previously reduced by means of layout techniques (e.g. regular layout designs).

The circuit used as variability sensor is the VCDL-based delay sensor described in Section 3.5, which contains a VCDL and a phase detector. An externally generated reference pulse passes through the VCDL which delays this pulse according to PVTA variations on that circuit region. The phase detector, consisting in a FF, generates a charge pulse as wide as the timing difference between the reference pulse falling edge and delayed pattern rising edge (see Figure 6.3). Finally, the VCDL end-to-end delay is locked to reference pulse T_{High} after around 2 μs.

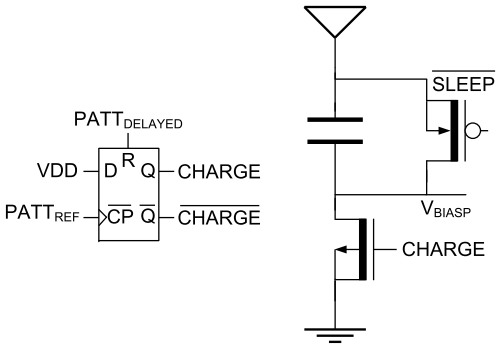

Figure 6.2 Phase detector (left) and charge pump (right) of the proposed FBB Generator.

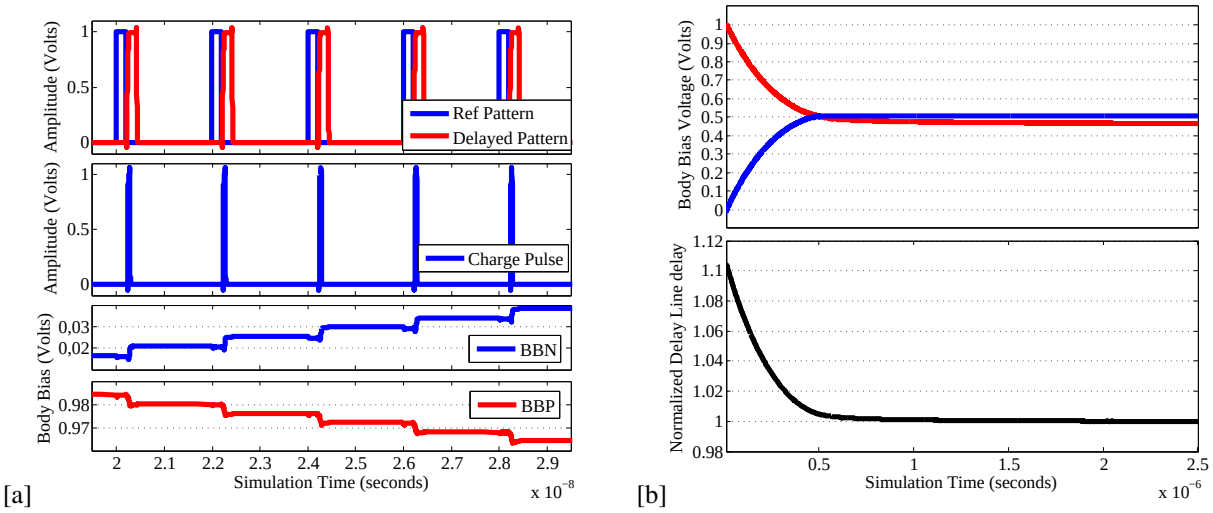

Figure 6.3 Phase detector control signals and charge pump output body biases of the FBB Generator (zoom) [a]. PMOS and NMOS generated body biases and VCDL end-to-end delay of the FBB Generator [b].

Figure 6.2 (right) shows the FBB generator. The first control transistor is used to discharge the capacitor when the body bias island is in sleep mode, and thus avoiding an increase in leakage power consumption during idle periods. This transistor W/L ratio is large enough to discharge the capacitor in 2 ns. The second control transistor converts the output pulse of the phase detector into a short current pulse that partially charges the capacitor. In this case a small W/L ratio is preferred to achieve a finer body bias adjust at the expense of the maximum speed of the dynamic variations that the circuit will be able to compensate.

In order to test the effectiveness of this BBG circuit, 500 Monte Carlo simulations were performed using the foundry-provided variability data. In absence of any control mechanism, the expected result is a histogram where samples are spread around the nominal VCDL end-to-end delay. The main limitation of a FBB-only generator is the impossibility to make circuits slower. Therefore, those circuits faster than the nominal case will be unnecessarily faster, with a consequent power consumption penalty. To overcome this issue, power supply voltage was decreased by 4% in order to make all the delay line samples slower than the nominal circuit simulation.

Figure 6.4 shows the VCDL end-to-end delay histograms for the 500 Monte Carlo simulations. As can be observed, before applying any adjustment (blue) delay samples are dispersed around the nominal delay with a coefficient of variation (understood as $3\sigma/\mu$) of 7%. When supply voltage is reduced by 4% normalized end-to-

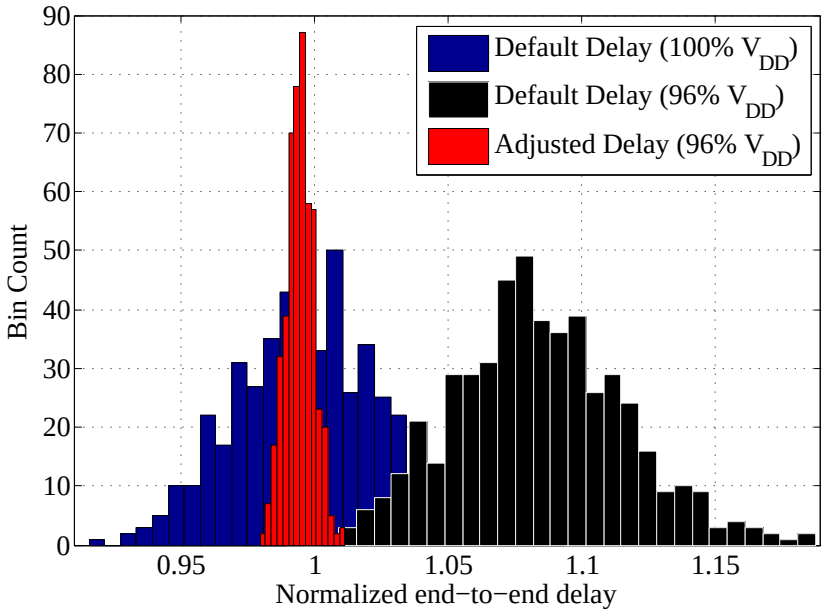

Figure 6.4 VCDL end-to-end delay histograms (a.u.) with and without applying FBB voltage.

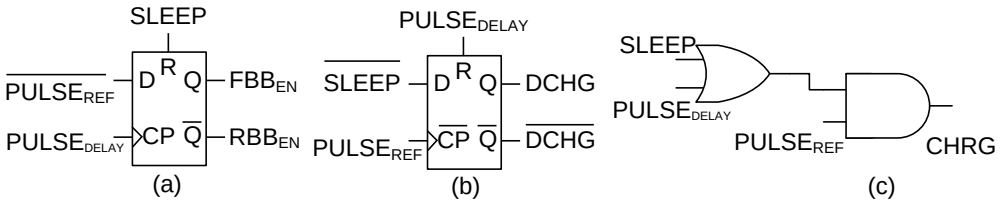

Figure 6.5 Phase detector of the proposed FBB+RBB generator.

end delay is shifted (black) and thus all delay samples are slower than the nominal case. If now the FBB generator is used to compensate variability (red), not only end-to-end delay of all the samples is dispersed around the nominal value, but also coefficient of variation decreases from 7% to 1%. What is more, dynamic power consumption is reduced by 8% as a side effect of having reduced power supply voltage.

6.2 FBB+RBB Generator circuit proposal

This second circuit evolves from the first proposed BBG. The objective was to extend post-silicon tuning range by implementing RBB capability. The major handicap of RBB is the out-of-rail voltage generation which makes this circuit more complex than its predecessor. The area penalty is also four times larger (346 μm^2).

The working principle is the same as the FBB generator since it also tracks VCDL end-to-end variations. The first difference is phase detector (see Figure 6.5), which generates both charge (RBB mode) and discharge (FBB mode) pulses as well as two signals that determine the BBG operation mode (RBB_{EN} or FBB_{EN}). Notice that the pulse width of charge (*CHRG*) and discharge (*DCHG*) signals will increase with the phase shift between nominal delay and current delay. Additionally, there is a *SLEEP* flag that forces the maximum RBB during idle periods. This circuit is described more in detail in Appendix C [58].

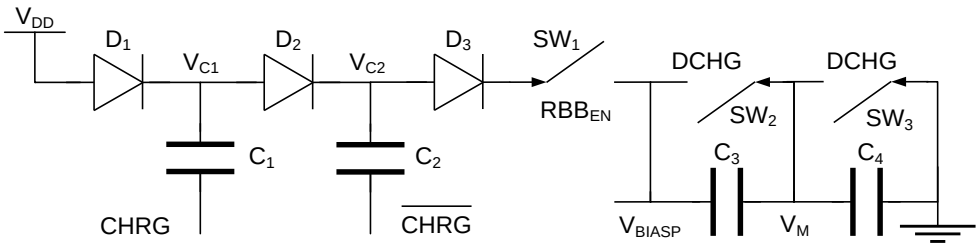

Figure 6.6 The proposed FBB+RBB Generator implements a charge pump based in a Dickson Charge Pump.

The charge pump design of the FBB+RBB also increases its complexity in comparison with the first proposal. Figure 6.6 shows the proposed charge pump inspired in a two-stage Dickson charge pump [81]. This kind of charge pump is commonly used as voltage multipliers and negative voltage supplies. In this proposal this circuit implementation is used to provide RBB voltages higher (for PMOS) and lower (for NMOS) than power supply voltage rails if needed. In addition to this feature, this circuit is also capable of generating FBB.

The role of C1 and C2 capacitors is to perform an AC coupling proportional to the *CHRG* pulse width. This overvoltage is transmitted to the output capacitors C3 and C4 only in RBB mode. Otherwise, FBB is produced by discharging C3 and C4 capacitors with *DCHG* signal. Further implementation details can be found in Appendix C [58].

As observed in waveforms of Figure 6.7, at the beginning of the simulation the circuit is forced to enter the sleep state for 2 μs and RBB mode is enabled whatever the output of the phase detector is. During sleep state the pattern pulse generator is used to charge the Dickson charge pump so that BBP and BBN get almost $2V_{DD}$ and $-V_{DD}$ respectively and, as a consequence, the BBI circuit including the VCDL transistors increase their V_{TH}. This increase causes a reduction in leakage power consumption by a factor of 3.9 according to the simulations. It also produces a dramatic increase in the end-to-end delay, which is not relevant during sleep states.

Once the sleep state stops, the circuit tracks the input pattern pulse by discharging capacitors C3 and C4 (see Figure 6.6) and thus reducing BBP and BBN voltages until the end-to-end delay of the VCDL is equal to pattern generator pulse width (steady state). This occurs around 5 μs after sleep mode has been disabled.

To test the effectiveness of this BBG, 300 Monte Carlo simulations were performed. In Figure 6.8 the normalized end-to-end delay of the VCDL before (in blue) and after (in red) applying body bias can be seen. Before applying any adjustment delay samples are dispersed around the nominal delay with a coefficient of variation of 12.38%. After FBB+RBB is applied, the coefficient of variation is reduced to 1.83%. This residual variability can be explained by two main reasons. First, phase detector logic also suffers from variability, and therefore the same phase shifts between input signals will produce marginally different outputs. Second, dead zone effect in the phase detector produces an overall system phase inaccuracy.

6.2.1 BBI Test

Once seen how the BBG is able to recover the nominal end-to-end delay of the internal VCDL, the next step is to evaluate its effectiveness as delay variability indicator. The test consists in reproducing an entire BBI, i.e. the circuit to protect against variability and the BBG itself, and then measure the slack time of the worst paths once the internal VCDL of the BBG is in steady state (locked). Slack time is defined as the maximum delay that combinational logic could accept without producing clock timing violations. In other words, it is the time margin between the worst path delay and next clock rising edge. To ensure that the circuit is working properly, the slack time must be greater than zero.

The sample circuit that we chose to protect against variability is a Cascaded Integrator Comb (CIC) filter with 8 coefficients since it is a widely used circuit in digital signal processing, it has an easy implementation, and worst paths are also easy to identify. The circuit simultaneously performs the sum of eight 16-bit input vectors in a single clock cycle and thus generating a 19-bit output vector (an extension of 3 bits is required to allow overflow).

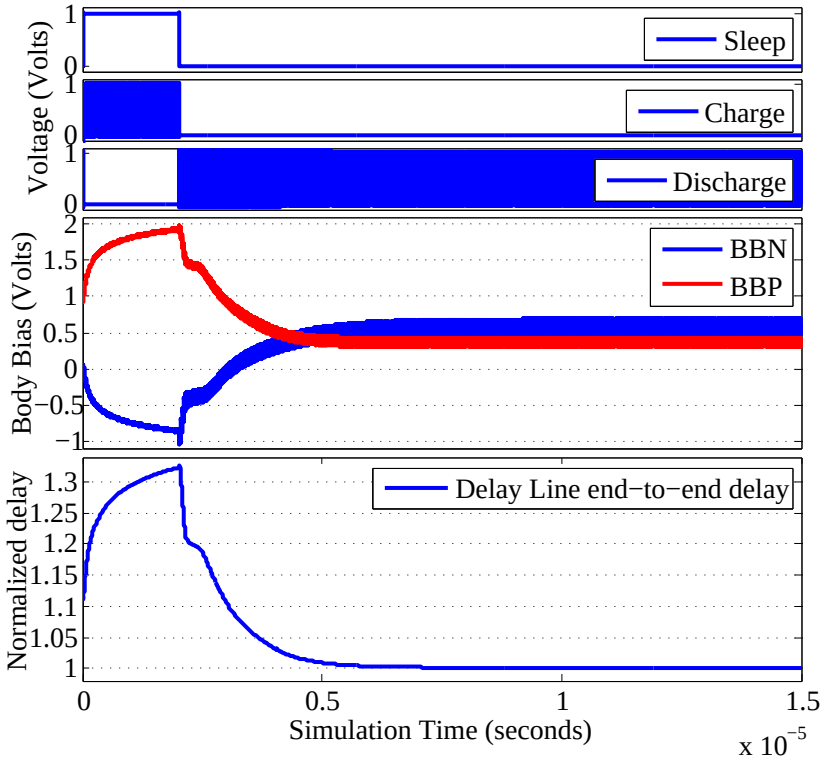

Figure 6.7 Phase detector control signals (top), PMOS and NMOS body biases (middle) and VCDL end-to-end delay of the FBB+RBB Generator.

The BBI testbench works as follows (see Figure 6.9): a functional block written in VerilogAMS generates those CIC filter input patterns which activate critical paths. As target clock frequency is set to 2.4 GHz, carry propagation delay should be equal or less than 416 ps. 600 ps later, patterns are input to the CIC filter, the functional block measures the elapsed time between new pattern is generated and the output value of the CIC filter is stable. To obtain the slack time from this measurement the measured value is subtracted from the clock period (416 ps). Negative slack measurements will indicate clock violations while positive measurements will mean that the circuit fulfills timing requirements. This measurement as well as the current VCDL end-to-end delay and body bias voltages (PMOS and NMOS) are stored in a measurement file for further analysis. Finally, input patterns are reset to 0 and after 1400 ps new input patterns are tested. This procedure is done every 2 ns until simulation ends after 7 μs, which is time enough for the VCDL to be locked according to Figure 6.7.

In parallel with the pattern generator block and CIC filter the proposed BBG with FBB+RBB capability adjusts PMOS and NMOS body biases so that VCDL recovers the nominal end-to-end delay. The frequency of the pulse pattern generator is 500 MHz, so that phase detectors of the BBG will generate charge or discharge pulses every 2 ns.

Therefore, each simulation generates a 3500-row CSV file corresponding to the body bias voltage measurements as well as VCDL end-to-end delay and CIC slack time. From these values we can extract how CIC slack time evolves as VCDL is locked to the nominal value and compare this measurement with the default slack time under ZBB (no body bias).

BBI was simulated only 26 times using the Monte Carlo method due to the fact that simulations required lots of computing resources. Figure 6.10 (a) shows how the slack time of CIC filter is improved by a factor of 3 when BBG is applied in contrast with ZBB. However, the coefficient of variation of slack time (3%) is clearly higher than the one for the VCDL which, according to simulation results depicted in Figure 6.8, is 1.83%. This delay

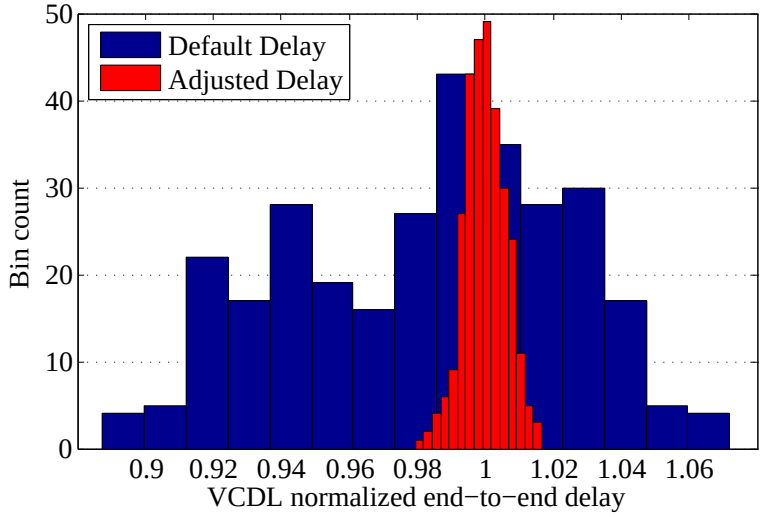

Figure 6.8 VCDL end-to-end delay histograms (a.u.) with and without applying FBB+RBB voltage (300 Monte Carlo samples).

variability worsening is mainly caused by mismatch, since delay variations produced by this phenomena are not spatially correlated, and hence VCDL cannot be sensitive to this effect.

The proposed BBG circuits work under the assumption that the monitoring circuit (VCDL) maintains a linear relationship with the circuit to protect against variability (CIC filter in this case) within the same BBI, i.e. VCDL end-to-end delay is proportional to CIC filter critical path delays. The effect of mismatch is, however, an unexpected delay offset in this linear relationship. To show this effect, slope and offset values are computed for each Monte Carlo sample by extracting coefficients of linear regression between VCDL end-to-end delay and CIC filter slack time measurements. Figure 6.10 (b) shows the linear relationship between VCDL and CIC for each Monte Carlo sample. It can be seen that slope coefficients do not vary substantially between samples, while offset vary in a range between ± 13 ps around the average value. Therefore, this offset in the relationship between the monitoring circuit and the circuit in the BBI produces an increase in timing variability, and thus a loss of effectiveness of proposed BBGs.

6.2.2 Mismatch reduction alternatives in BBIs

Despite the fact that proposed BBGs noticeably reduce the effects of variability, these circuits are not able to correct mismatch, but nevertheless some improvements could be done in order to maximize Parametric Yield:

- **Increase design margin:** this solution is acceptable since most of the variability is already mitigated, and hence 2% of clock frequency reduction would be enough to maximize Parametric Yield.
- **Combine BBG with regular layouts:** demonstrator chip measurements summarized in Chapter 5 showed that the most of the mismatch is caused by systematic causes, and regular layouts were proved to be effective in reducing such kind of variations.
- **Critical path(s) replica:** some authors propose the use of replicas of critical path(s) instead of VCDLs [14][18], since mismatch sensitivity depends on transistor dimensions, and therefore, identical critical paths will suffer less mismatch in comparison with a generic VCDL. The main drawback of this improvement proposal is that BBG design is no longer equal for each BBI, which will make design automation more difficult.
- **VCDL calibration:** there is also a post-silicon alternative that consists in calibrating the VCDL during the Power-On Self-Test (POST) of the IC. During this test period each BBI would activate their critical paths and

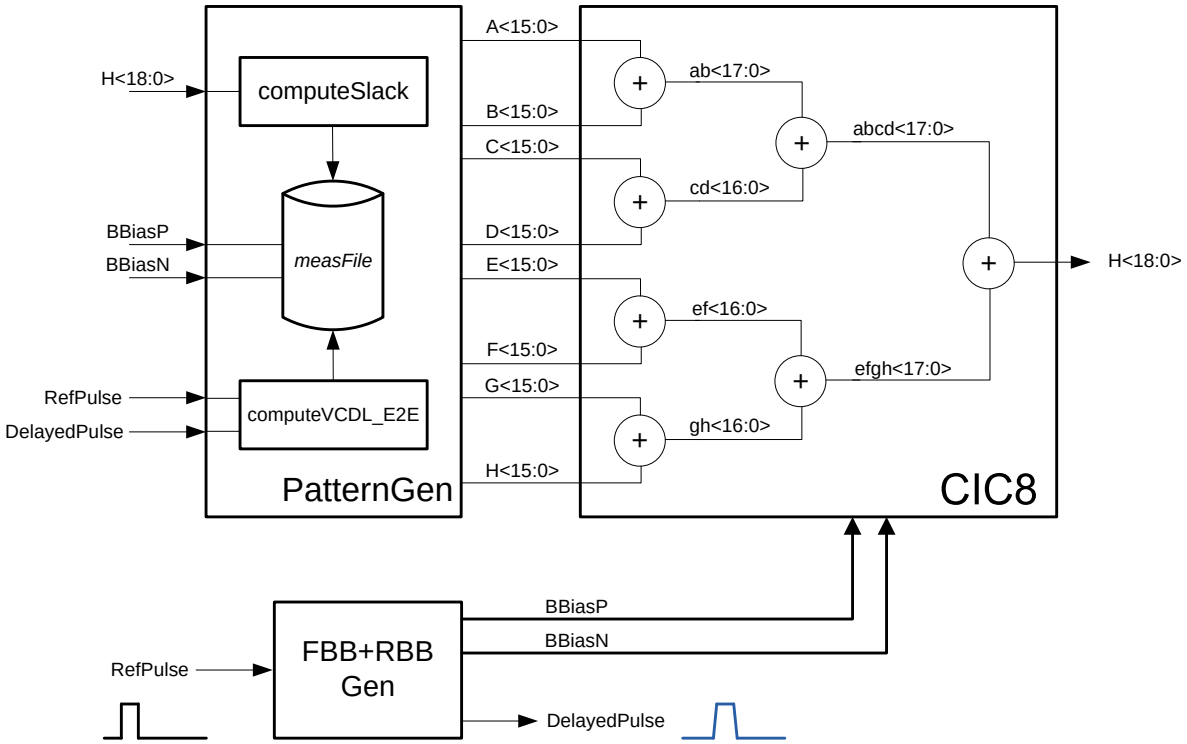

Figure 6.9 Block diagram of the BBI testbench.

VCDLs would decrease their length (in number of stages) until a clock violation occurs. Once the optimal VCDL length is found, the BBG is ready to track dynamic variations along BBI. This functionality would require a tunable length VCDL, and the insertion of scan chains to set and observe the signals across critical paths. The main advantage is that no design margins or special layout designs are required. However, BBG complexity, which is one of the assets of our proposal, would substantially increase and thus area overhead.

6.3 Comparison with other proposals

The area penalty of both proposals were estimated (layouts were not implemented) using the area of each component with a standard cell library for logic elements (phase detector and VCDL) and doing a tentative layout placement for capacitors. The main contribution to the area penalty are the charge pump capacitances. Capacitors for the FBB-only proposal occupy $64\mu m^2$, while capacitors for the FBB+RBB proposal occupy 320 μm^2. As for transistor resources, FBB-only and FBB+RBB proposals require 17 μm^2 and 26 μm^2 respectively.

Power consumption overhead was also been computed. As this power consumption depends on the external pattern generator frequency, simulations were done for different pattern frequencies and thus obtaining dynamic power consumption as a function of pattern frequency. As expected, dynamic power consumption overhead exhibits a linear dependency with the pattern frequency, being 33 and 38 μW per GHz for FBB-only and FBB+RBB proposals respectively.

A comparison chart of area overhead is presented in Table 6.2. This table shows a comparison between the area of the proposed BBG in this article, and similar proposals in literature. In order to make a fair comparison between different technologies, circuit areas have been normalized (see Normalized area row) by dividing them by the square of the technology node and then, again, normalized to our circuit. Therefore, even assuming that circuits from other proposals would scale proportionally, our BBG proposals are significantly smaller. The main

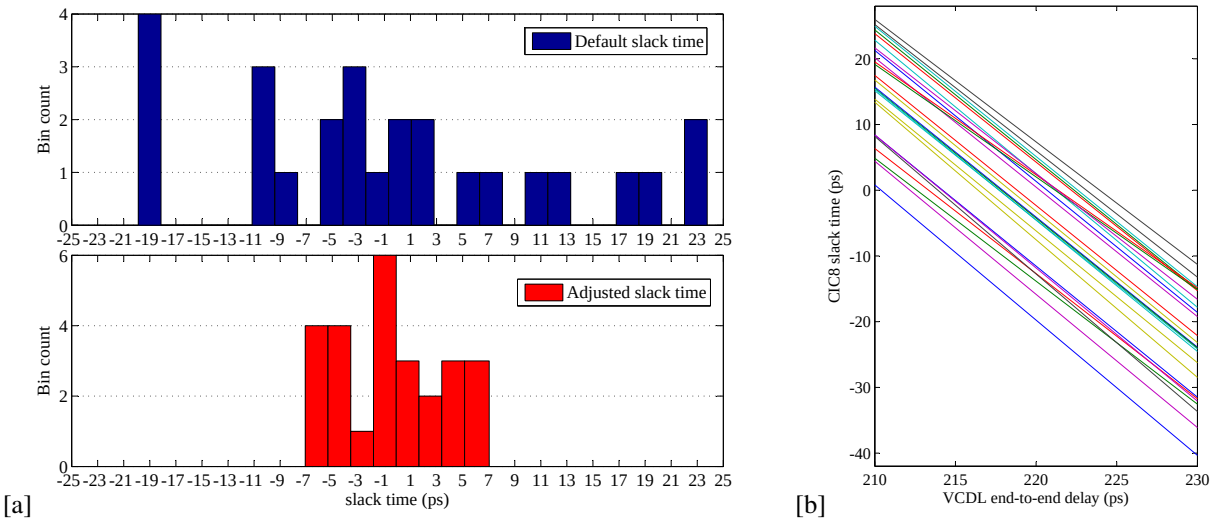

Figure 6.10 CIC slack time histograms with and without applying FBB+RBB voltage (26 Monte Carlo samples) [a]. Linear relationship between VCDL end-to-end delay and CIC filter slack time [b].

reason that explains this area improvement is the use of a VCDL in closed loop instead of DACs to generate body bias.

	FBB Proposal	**FBB+RBB Proposal**	**[68]**	**[15]**
Function	FBB	FBB+RBB	FBB+RBB	FBB
Process	28 nm	28 nm	65 nm	90 nm
Circuit area	81 μm^2	346 μm^2	5200 μm^2	30000 μm^2
Normalized area	0.23X	1X	2.79X	8.39X
Power consumption	33 $\mu W/GHz$	38 $\mu W/GHz$	600 μW	210 μW

Table 6.2 Comparison of the proposed BBGs with other proposals.

CHAPTER 7

CONCLUSIONS & FUTURE WORK

Since the decade of the 1960's microelectronics has advanced by leaps and bounds due to continuous transistor scaling. In 40 years semiconductor industry has gone from integrating thousands of 10 μm long transistors to integrate thousands of millions of only a few nm. Scale productions as well as a sustained increasing demand have democratized consumer electronics. However, silicon transistor scaling is threatened at medium term by the fact that transistor sizes are getting dimensions comparable with atomic structures of materials involved in manufacturing process, and lithography wavelength hold back. All these handicaps triggered an increment of variability and an increase in faulty dies per wafer, and thus questioning the suitability of continuous transistor scaling. The main objective of this thesis is to find the optimal design of an adaptive system to combat PVTA variations.

The most relevant milestone in this thesis is the minimization of variability effects by implementing a body bias generator which automatically adapts to local chip variations. In this way transistor delay variation coefficient is reduced from 9% to 3%. Its simplicity as well as the reduced dimensions and low power consumption makes this proposal unique.

As further explained, transistor delay was chosen as variability indicator due to being a parameter which has a good sensitivity to all the types of PVTA variations that might appear. Despite this, it is impossible to completely mitigate variability effects in chip performance and power consumption, since it would imply to tune each transistor independently which is clearly infeasible. Nevertheless, the use of adaptive techniques allows us to noticeably improve Parametric Yield.

It is worth mentioning that there were some variations in the initial thesis approach as for the variability mitigation circuit proposal. Some changes, such as discard the use of AVS or opt for a VCDL-based delay sensor instead of a TDC-based one, have been motivated by the demonstrator chip measurement results that suggested to increase the number of circuit islands as possible and thus treating variability at more local level. On the other hand, the irruption of FD-SOI technologies, which enable a wide range of body bias voltages, gives a second youth to ABB technique which was getting deprecated due to the low adjust range that provided in submicronic bulk CMOS technologies.

Another not less important reached goal has been the contribution to knowledge by means of results that empirically demonstrates that regular layouts are less affected by WID variations. At chip level, it can be seen that conventional (non-regular) layouts suffer large differences in circuit performance according to the physical location within the chip, while this differences are reduced substantially for the case of regular layouts. At transistor level, the effect of systematic variations is almost inappreciable in regular layouts and thus reducing transistor mismatch to almost intrinsic random variations. Conversely, systematic variations in non-regular layouts produce an increase in mismatch ranged between 2.5 and 4. Moreover, comparing simulated and measured transistor delay variability we conclude that simulator is not able to predict the improvements in transistor mismatch brought by layout techniques.

Despite it is said that the number of dopants will be one of the most critical aspects as for process variability in sub-20 nm technologies [5], both 65 nm simulations and 40 nm chip measurements still point to transistor channel length as the most important parameter.

7.1 Variability indicator choice

One of the main objectives of this thesis is to determine which observable magnitude gives better information about PVTA variations suffered by ICs. In practice this means that the magnitude used as variability indicator should be sensitive to any of the variability phenomena that may occur to an integrated circuit.

According to simulations, transistor delay is the only observable magnitude that has an acceptable sensitivity to PVT variations. On the other hand, and despite that static power consumption is also sensitive to PVT variations, the huge dependency with process variations masks voltage and temperature variations. Notice that ageing simulations could not be done since transistor models did not support this feature. Nevertheless, it is well known that transistor ageing increases its V_{TH} and thus increasing combinational logic delay.

Moreover, the use of transistor delay as an observable magnitude is proved to be the most effective solution to reduce overall variability when post-silicon techniques (ABB and AVS) are simulated. Indeed, not only delay variability but also leakage variability is reduced. However, dynamic power consumption variability is slightly worsened while it essentially does not increase in average when using delay as the observable magnitude.

In parallel, several sensing strategies were studied. On the one hand, thermal coupling was proposed to monitor power consumption variability. Once characterized transistors as heat sources we concluded that the heat released by transistors due to leakage was too small as to be measured with an acceptable resolution, while thermal gradient produced by dynamic power was large enough as to be measured with good resolution. On the other hand, delay lines were proposed to monitor transistor delay sensors. After analysing pros and cons we concluded that TDC-based proposal would imply larger area and power consumption overheads which, in the end, would not make practical our sensing proposal. Therefore, we propose a circuit based in a VCDL that encodes transistor delay variations into a square pulse.

7.2 Variability control strategies

According to demonstrator chip measurements, the most of variations in conventional physical designs (i.e. non-regular layouts) are intra-die. This means that different parts of the IC are differently affected by process variations. The only way to face this problem with adaptive techniques is by partitioning chips in islands or regions. What is more, as huge variations may occur in few hundreds of microns the number of islands should be a relatively large number (100 per square millimetre, for example). This makes infeasible the use of voltage scaling technique since each voltage island would require its own voltage regulator and this would produce and overwhelming power consumption and area overheads.

New semiconductor technologies UTBB FD-SOI provide a wide range of body bias voltages, more precisely between -1.8 to V_{DD}-0.3 volts. This wide range of voltages knobs the V_{TH} of transistors which, in turn, produce substantial changes in transistor delay, and hence makes attractive the use of ABB to control variability effects. From simulation results we conclude that transistor delay adjust range is ranged between -29.7% to +12.6% in respect to nominal value, wide enough as to compensate PVTA variations in most of the scenarios.

Two types of body bias generators were proposed. On the one hand, a BBG that only implements FBB, which means that the voltage range provided by such kind of generator is limited to values comprised within supply voltage rails, and not higher than 0.7 volts respect to default body bias (ZBB). In practise, this involves reducing transistor delay tuning range to positive values only, i.e. from 0 to +12.6%. Despite the simplicity of the design, the obtained results are quite satisfactory and demonstrate that it could be a good alternative for a low variability scenario given the low power consumption and area penalty of this proposal. On the other hand, and fruit of the evolution of the first BBG a second generator with FBB+RBB capability is proposed. The main virtue of this generator is the out-of-rail voltage capability which enables to fully exploit ABB technique. To achieve this goal a novel charge pump circuit is designed. This circuit consists of a modified Dickson's charge pump which provides positive and negative voltages according to the error phase measured by delay variability monitor. In this case the adjust range is wide enough as to cover high environmental variability scenarios.

Finally, we implement a testbench where FBB+RBB generator is assessed. In this testbench the BBG monitors (in a non-invasive way) the process variations that an adjacent block is suffering, and generates bias voltages in order to compensate the effects of variability of the adjacent block. Comparing variability of the BBG's internal VCDL and the adjacent block after the adaptive adjust, a slight worsening is observed which is mainly caused by transistor mismatch due to the lack of spatial correlation. Even so, the coefficient of variation of slack time is reduced by a factor of 3. In order to improve the obtained results several alternative are proposed, of which the clock frequency reduction by 2% is the most practical one.

As regards the Parametric Yield improvement which our proposal can contribute, it is worth to say that there are not enough data to give a reliable number, since simulation results are not statistically significant neither this proposal has been tested with other circuits apart from the CIC. Moreover, this Parametric Yield improvement would also depend on the tolerance margins set and the expected number of faulty chips when any adaptive technique is applied. Nevertheless, the observed trend in simulations points out that variability effects can be reduced by 66%. On the other hand, the placement of BBGs implies area penalty. Assuming that we set a density of 100 BBIs per square millimetre, i.e. BBIs of 100 μm^2, the area overhead of the FBB+RBB proposal would be around 3.5%. In this scenario, the Parametric Yield increase achieved by our post-silicon tuning proposal must be higher than 3.5% to compensate the area penalty and thus make our post-silicon tuning proposal profitable.

Comparing proposed BBGs in this thesis with other authors' it can be seen that our proposals substantially improve power consumption and area penalty. The main novelty and the reason for low overhead achievements lies in the fact that delay variability is directly encoded into an analog voltage, which not only enables to dispense with DACs but also simplifies transistor delay monitoring. Moreover, as each BBI works in closed loop, they do not require any external control and thus making this proposal scalable and easy to integrate into an automated EDA design flow.

7.3 Future Work

In this section some future research lines are outlined. These research lines not only would give continuity to the research done along this thesis, but they would also make this proposal interesting for the semiconductor industry.

The first step would be the implementation of a BBG library. On the one hand, a set of BBGs optimized to compensate a given dynamic variation bandwidth (caused by noise and interferences coupled into power supply rails) could be implemented. It is worth mentioning that higher bandwidths imply higher power consumption by the BBG, while low bandwidths require high capacitive values and thus area overhead. On the other hand, BBG layout must be designed following the same geometry rules as standard cells: cell height, supply voltage rails continuity, horizontal grid spacing, etc. Once layouts are implemented layout abstract views have to be done in order to enable the EDA tool to perform automated placement of designs. In order to limit the area penalty impact, which is mainly caused by charge pump capacitors, and as long as the process option is supported by technology and it is worth for money, capacitors could be implemented with Metal-in-Metal capacitors which are implemented in dedicated upper metal layers, thus reducing the area overhead between 80 and 90%.

In order to enable automated physical design of chips it will be mandatory to modify standard cells since body terminals of transistors are connected to power supply by default (ZBB). The easiest way to adapt cells to our needs is to split power supply stripes as shown in Figure 7.1.b, so that bulk terminals are connected to a dedicated

terminal (BBP or BBN). Notice that a new spacer filler cell is required to delimit contiguous BBIs (see Figure 7.1.c). Finally, BBG would be inserted in the middle of the island and the routing algorithm should interconnect PMOS and NMOS body bias voltages outputs of the generator with the horizontal body bias stripes as seen in the example depicted in Figure 7.2.

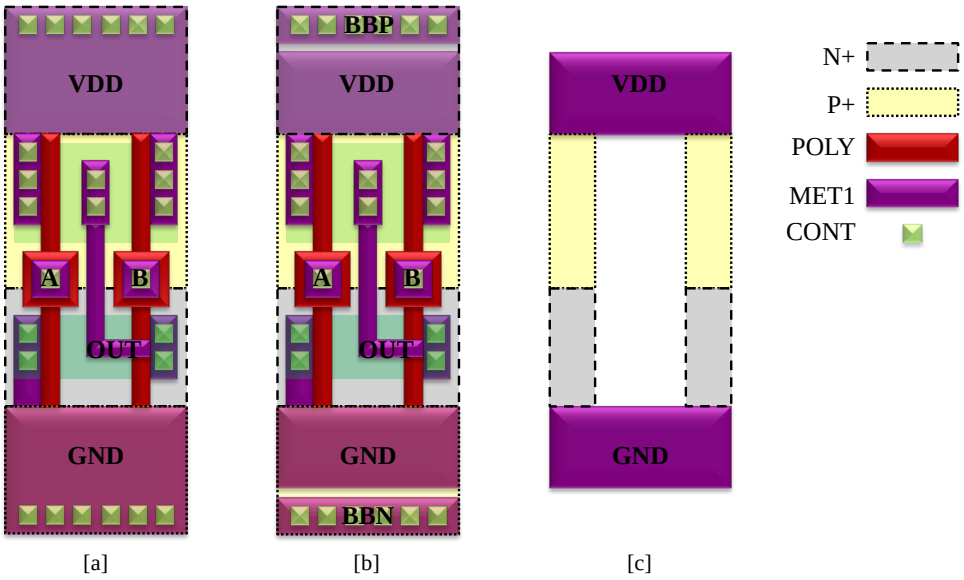

Figure 7.1 Orignal standard cell (NAND2) layout design [a]. Modified standard cell (NAND2) with dedicated body terminals [b]. Spacer cell to delimit contiguous BBIs [c].

Given the experience of HIPICS group in the design of regular layouts and given the effectiveness of such kind of designs for reducing process variability at design stage, another possible future line of research could be the implementation of a regular standard cell library and combining them with the FBB-only generator proposed in this thesis. Despite tuning range of this generator is clearly lower than the maximum enabled by technology, it would probably be enough since global variations suffered by regular designs are noticeably lower than non-regular designs. Moreover, the big advantage of this combination of techniques is that the offset between optimal body bias voltages for BBG and monitored circuit would decrease substantially since it is mainly caused by transistor mismatch as seen in Figure 6.10. Therefore, the coefficient of variation would decrease and, on the contrary, Parametric Yield would increase.

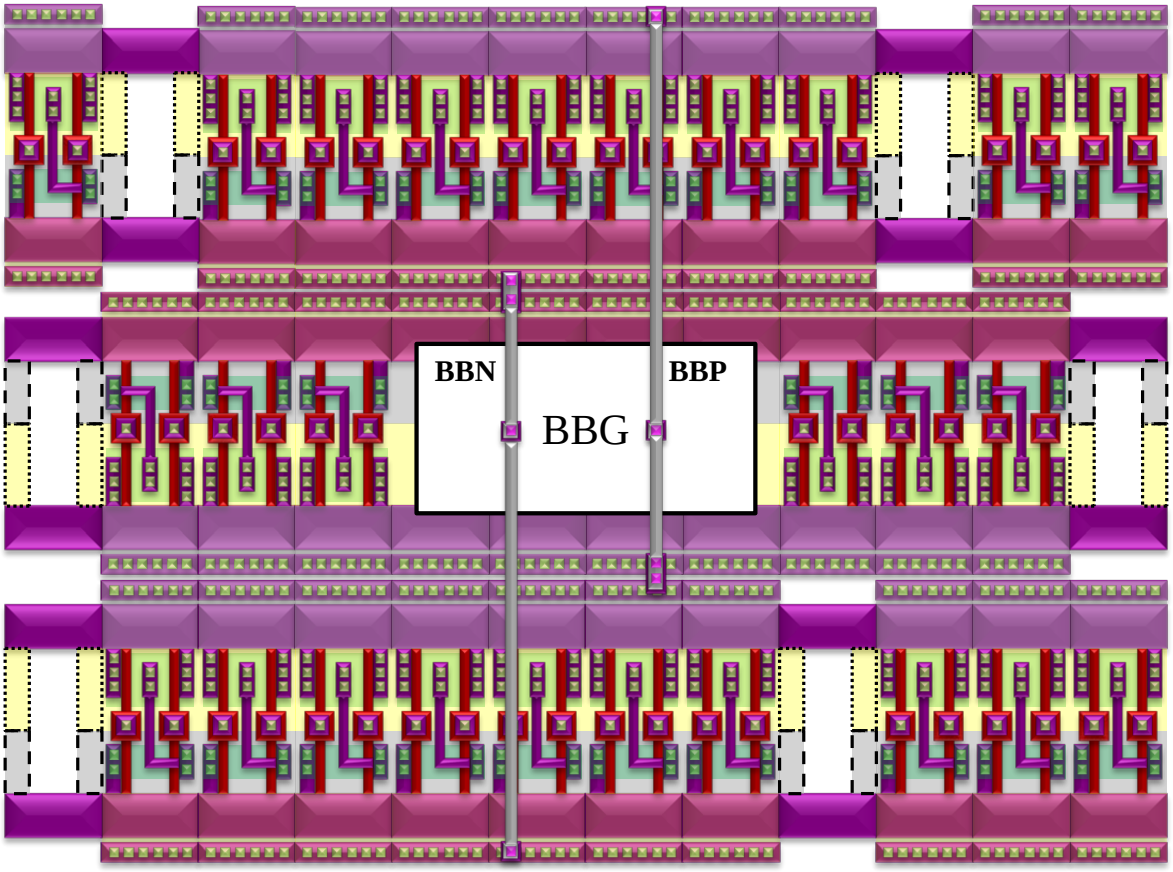

Figure 7.2 An example of a BBI. BBG is placed in the middle of the island and bias voltages, BBP and BBN, are connected to PMOS and NMOS body termianal rails respectively. Body bias delimiters isolates the island from neighbor islands.

REFERENCES

[1] Intel. (2005, April) Moores's law 40th anniversary. [Online]. Available: http://www.intel.com/pressroom/kits/events/moores_law_40th/

[2] Wikipedia. (2015, June) Transistor count. [Online]. Available: https://en.wikipedia.org/wiki/Transistor_count

[3] ——. (2015, June) List of intel microprocessors. [Online]. Available: https://en.wikipedia.org/wiki/List_of_Intel_microprocessors

[4] ——. (2015, June) Itanium. [Online]. Available: https://en.wikipedia.org/wiki/Itanium

[5] A. Asenov. (2010, July) Statistical variability in conventional and modern cmos devices. [Online]. Available: https://eel.postgrau.upc.edu/courses-and-seminars/courses-materials/2009-2010/barcelona1-07-10.pdf

[6] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov, "Understanding ler-induced statistical variability: A 35,000 sample 3d simulation study," in *Solid State Device Research Conference, 2009. ESSDERC '09. Proceedings of the European*, Sept 2009, pp. 423–426.

[7] S. Jilla. (2009, November) Confronting manufacturing closure at 32nm and below. [Online]. Available: http://www.eejournal.com/archives/articles/20091117-mentor

[8] M. Choi and L. Milor, "Impact on circuit performance of deterministic within-die variation in nanoscale semiconductor manufacturing," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 7, pp. 1350–1367, July 2006.

[9] S. Gómez and F. Moll, "Lithography Aware Regular Cell Design based on a Predictive Technology Model," *Journal of Low Power Electronics*, vol. 6, no. 4, pp. 588–600, 2010.

[10] K. Bowman. (2010, June) Variability in microprocessor logic design: Trends, sources, consequences, & solutions. [Online]. Available: http://eel.postgrau.upc.edu/research/courses-and-seminars/courses-materials/2009-2010/keith-bowman-sessioni.pdf

[11] Teklatech. (2013) Floordirector rail-aware optimization. [Online]. Available: http://teklatech.com/web2/rail-aware-optimization-2

[12] S.-W. Chen, M.-H. Chang, W.-C. Hsieh, and W. Hwang, "Fully on-chip temperature, process, and voltage sensors," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, May 2010, pp. 897–900.

[13] N. Azizi and F. Najm, "Compensation for within-die variations in dynamic logic by using body-bias," in *IEEE-NEWCAS Conference, 2005. The 3rd International*, June 2005, pp. 167–170.

[14] R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "Mitigating parameter variation with dynamic fine-grain body biasing," in *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*, Dec 2007, pp. 27–42.

[15] M. Meijer, J. 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 *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, May 2010, pp. 2482–2485.

[16] N. Kamae, A. Tsuchiya, and H. Onodera, "An area effective forward/reverse body bias generator for within-die variability compensation," in *Solid State Circuits Conference (A-SSCC), 2011 IEEE Asian*, Nov 2011, pp. 217–220.

[17] M. Meijer, F. Pessolano, and J. de Gyvez, "Limits to performance spread tuning using adaptive voltage and body biasing," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, May 2005, pp. 5–8 Vol. 1.

[18] A. K. Uht, "Going beyond worst-case specs with teatime," *Computer*, vol. 37, no. 3, pp. 51–56, 2004.

[19] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. Harris, D. Blaauw, and D. Sylvester, "Bubble razor: An architecture-independent approach to timing-error detection and correction," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb 2012, pp. 488–490.

[20] E. Beigne and P. Vivet, "An innovative local adaptive voltage scaling architecture for on-chip variability compensation," in *New Circuits and Systems Conference (NEWCAS), 2011 IEEE 9th International*, June 2011, pp. 510–513.

[21] K. Bowman, J. Tschanz, S. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik, and V. De, "A 45 nm resilient microprocessor core for dynamic variation tolerance," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 194–208, Jan 2011.

[22] S. Nassif, K. Bernstein, D. Frank, A. Gattiker, W. Haensch, B. Ji, E. Nowak, D. Pearson, and N. Rohrer, "High performance cmos variability in the 65nm regime and beyond," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, Dec 2007, pp. 569–571.

[23] L. T. Pang, "Measurement and analysis of variability in cmos circuits," Ph.D. dissertation, EECS Department, University of California, Berkeley, Aug 2008. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-108.html

[24] J. Soden and C. Hawkins, "Iddq testing: issues present and future," *Design Test of Computers, IEEE*, vol. 13, no. 4, pp. 61–65, Winter 1996.

[25] B. P. Wong, A. Mittal, G. W. Starr, F. Zach, V. Moroz, and A. Kahng, *Nano-CMOS Design for Manufacturability: Robust Circuit and Physical Design for Sub-65Nm Technology Nodes*. New York, NY, USA: Wiley-Interscience, 2008.

[26] T. Chen and S. Naffziger, "Comparison of adaptive body bias (abb) and adaptive supply voltage (asv) for improving delay and leakage under the presence of process variation," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 11, no. 5, pp. 888–899, Oct. 2003. [Online]. Available: http://dx.doi.org/10.1109/TVLSI.2003.817120

[27] G. E. Moore, "Progress in digital integrated electronics," *IEDM Tech Digest*, pp. 11–13, 1975.

[28] ——, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, April 1965.

[29] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits*. Prentice hall Englewood Cliffs, 2002, vol. 2.

[30] J. Altet, W. Claeys, S. Dilhaire, and A. Rubio, "Dynamic surface temperature measurements in ics," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1519–1533, Aug 2006.

[31] R. Swanson and J. Meindl, "Ion-implanted complementary mos transistors in low-voltage circuits," *Solid-State Circuits, IEEE Journal of*, vol. 7, no. 2, pp. 146–153, Apr 1972.

[32] J. Kao, S. Narendra, and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques [ic cad tools]," in *Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on*, Nov 2002, pp. 141–148.

[33] S. Chandra, K. Lahiri, A. Raghunathan, and S. Dey, "Variation-tolerant dynamic power management at the system-level," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 9, pp. 1220–1232, Sept 2009.

[34] A. Kumar, L. Shang, L.-S. Peh, and N. Jha, "System-level dynamic thermal management for high-performance microprocessors," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 27, no. 1, pp. 96–108, Jan 2008.

[35] J. Figueras and E. Macii, "Leakage power modeling and control strategies," Sept 2008, slides of a talk given at Clean Workshop, September 8, Lisbon, Portugal.

[36] S. Garg and D. Marculescu, "System-level leakage variability mitigation for mpsoc platforms using body-bias islands," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 12, pp. 2289–2301, Dec 2012.

[37] M. Pons, F. Moll, A. Rubio, J. Abella, X. Vera, and A. Gonzalez, "Vcta: A via-configurable transistor array regular fabric," in *VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP*, Sept 2010, pp. 335–340.

[38] A. Asenov, "Random Dopant Induced Threshold Voltage Lowering and Fluctuations in Sub-0.1 (micron)meter MOSFET's: A 3-D 'Atomistic' Simulation Study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505–2513, 1998.

[39] W. Tiefeng, Z. HeMing, and H. HuiYong, "Research on the static characteristics of cmos circuits in the effects of gate tunneling current," in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, vol. 3, May 2010, pp. 173–177.

[40] J. Cai and C.-T. Sah, "Gate tunneling currents in ultrathin oxide metaloxidesilicon transistors," *Journal of Applied Physics*, vol. 89, no. 4, 2001.

[41] A. Nandi, S. Jha, and A. Kumari, "Effect of some device parameters on the transient characteristics of nanoscale cmos inverters," in *Devices, Circuits and Communications (ICDCCom), 2014 International Conference on*, Sept 2014, pp. 1–6.

[42] E. Ungersboeck, S. Dhar, G. Karlowatz, V. Sverdlov, H. Kosina, and S. Selberherr, "The effect of general strain on the band structure and electron mobility of silicon," *Electron Devices, IEEE Transactions on*, vol. 54, no. 9, pp. 2183–2190, Sept 2007.

[43] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 11, pp. 1396–1402, Nov 2002.

[44] T. Chawla, S. Marchal, A. Amara, and A. Vladimirescu, "Impact of intra-die random variations on clock tree," in *NORCHIP, 2009*, Nov 2009, pp. 1–4.

[45] D. Reid, C. Millar, S. Roy, and A. Asenov, "Statistical enhancement of the evaluation of combined rdd- and ler-induced v_T variability: Lessons from 10^5 sample simulations," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2257–2265, Aug 2011.

[46] J. Lee, M. Rotaru, M. Iyer, H. Kim, and J. Kim, "Analysis and suppression of ssn noise coupling between power/ground plane cavities through cutouts in multilayer packages and pcbs," *Advanced Packaging, IEEE Transactions on*, vol. 28, no. 2, pp. 298–309, May 2005.

[47] Y.-L. Chuang, P.-W. Lee, and Y.-W. Chang, "Voltage-drop aware analytical placement by global power spreading for mixed-size circuit designs," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, Nov 2009, pp. 666–673.

[48] T.-H. Ding and Y.-S. Li, "Efficient method for modeling of ssn using time-domain impedance function and noise suppression analysis," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 2, no. 3, pp. 510–520, March 2012.

[49] A. Chakraborty and D. Pan, "Controlling nbti degradation during static burn-in testing," in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, Jan 2011, pp. 597–602.

[50] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, and M. Nelhiebel, "Recent advances in understanding the bias temperature instability," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, Dec 2010, pp. 4.4.1–4.4.4.

[51] B. Tenbroek, M. Lee, W. Redman-White, R. Bunyan, and M. Uren, "Impact of self-heating and thermal coupling on analog circuits in soi cmos," *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 7, pp. 1037–1046, Jul 1998.

[52] R. Zheng, J. Velamala, V. Reddy, V. Balakrishnan, E. Mintarno, S. Mitra, S. Krishnan, and Y. Cao, "Circuit aging prediction for low-power operation," in *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*, Sept 2009, pp. 427–430.

[53] S. Han, J. Choung, B.-S. Kim, B. H. Lee, H. Choi, and J. Kim, "Statistical aging analysis with process variation consideration," in *Computer-Aided Design (ICCAD), 2011 IEEE/ACM International Conference on*, Nov 2011, pp. 412–419.

[54] Y. Wang, H. Luo, K. He, R. Luo, H. Yang, and Y. Xie, "Temperature-aware nbti modeling and the impact of input vector control on performance degradation," in *Design, Automation Test in Europe Conference Exhibition, 2007. DATE '07*, April 2007, pp. 1–6.

[55] C. Tan, *Electromigration in Ulsi Interconnections*, ser. International Series on Advances in Solid State Electronics & Technology. World Scientific, 2010. [Online]. Available: https://books.google.es/books?id=1icQpBUnQD4C

[56] J. Mauricio, F. Moll, and J. Altet, "Monitor strategies for variability reduction considering correlation between power and timing variability," in *SOC Conference (SOCC), 2011 IEEE International*, Sept 2011, pp. 225–230.

[57] J. Mauricio, F. Moll, and S. Gomez, "Measurements of process variability in 40-nm regular and nonregular layouts," *Electron Devices, IEEE Transactions on*, vol. 61, no. 2, pp. 365–371, Feb 2014.

[58] J. Mauricio and F. Moll, "Local variations compensation with dll-based body bias generator for utbb fd-soi technology," in *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, June 2015.

[59] H. Balachandran, K. Butler, and N. Simpson, "Facilitating rapid first silicon debug," in *Test Conference, 2002. Proceedings. International*, 2002, pp. 628–637.

[60] J. Altet, A. Rubio, and H. Tamamoto, "Analysis of the feasibility of dynamic thermal testing in digital circuits," in *Test Symposium, 1997. (ATS '97) Proceedings., Sixth Asian*, Nov 1997, pp. 149–154.

[61] Y. Zhang, A. Srivastava, and M. Zahran, "Chip level thermal profile estimation using on-chip temperature sensors," in *Computer Design, 2008. ICCD 2008. IEEE International Conference on*, Oct 2008, pp. 432–437.

[62] S. Ghosh, S. Bhunia, A. Raychowdhury, and K. Roy, "A novel delay fault testing methodology using low-overhead built-in delay sensor," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 12, pp. 2934–2943, Dec 2006.

[63] M.-C. Tsai, C.-H. Cheng, and C.-M. Yang, "An all-digital high-precision built-in delay time measurement circuit," in *VLSI Test Symposium, 2008. VTS 2008. 26th IEEE*, April 2008, pp. 249–254.

[64] X. Wang, M. Tehranipoor, and R. Datta, "A novel architecture for on-chip path delay measurement," in *Test Conference, 2009. ITC 2009. International*, Nov 2009, pp. 1–10.

[65] M. Farahat, F. Farag, and H. Elsimary, "Only digital technology analog-to-digital converter circuit," in *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*, vol. 1, Dec 2003, pp. 178–181 Vol. 1.

[66] S. Garg and D. Marculescu, "System-level mitigation of WID leakage power variability using body-bias islands," in *Proceedings of the 6th IEEE/ACM/IFIP international conference on Hardware/Software codesign and system synthesis - CODES/ISSS '08*. New York, New York, USA: ACM Press, Jan. 2008, p. 273. [Online]. Available: http://www.researchgate.net/publication/221656795_System-level_mitigation_of_WID_leakage_power_variability_using_body-bias_islands

[67] N. Kamae, A. Tsuchiya, and H. Onodera, "A body bias generator compatible with cell-based design flow for within-die variability compensation," in *Solid State Circuits Conference (A-SSCC), 2012 IEEE Asian*, Nov 2012, pp. 389–392.

[68] N. Kamae, A. Islam, A. Tsuchiya, and H. Onodera, "A body bias generator with wide supply-range down to threshold voltage for within-die variability compensation," in *Solid-State Circuits Conference (A-SSCC), 2014 IEEE Asian*, Nov 2014, pp. 53–56.

[69] T. Kuroda, K. Suzuki, S. Mita, T. Fujita, F. Yamane, F. Sano, A. Chiba, Y. Watanabe, K. Matsuda, T. Maeda, T. Sakurai, and T. Furuyama, "Variable supply-voltage scheme for low-power high-speed cmos digital design," *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 3, pp. 454–462, Mar 1998.

[70] K. Nowka, G. Carpenter, E. MacDonald, H. Ngo, B. Brock, K. Ishii, T. Nguyen, and J. Burns, "A 32-bit powerpc system-on-a-chip with support for dynamic voltage scaling and dynamic frequency scaling," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 11, pp. 1441–1447, Nov 2002.

[71] J. Lee, S.-T. Zhou, and N. S. Kim, "Analyzing impact of multiple abb and avs domains on throughput of power and thermal-constrained multi-core processors," in *Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific*, Jan 2010, pp. 229–234.

[72] H. Mostafa, M. Anis, and M. Elmasry, "A novel low area overhead direct adaptive body bias (d-abb) circuit for die-to-die and within-die variations compensation," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 10, pp. 1848–1860, Oct 2011.

[73] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: a low-power pipeline based on circuit-level timing speculation," in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, Dec 2003, pp. 7–18.

[74] K. Bowman, J. Tschanz, N. S. Kim, J. Lee, C. Wilkerson, S. Lu, T. Karnik, and V. De, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 49–63, Jan 2009.

[75] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," in *Low Power Electronics and Design, 2002. ISLPED '02. Proceedings of the 2002 International Symposium on*, 2002, pp. 64–67.

[76] J. Altet and A. Rubio, *Thermal Testing of Integrated Circuits*. Springer-Verlag New York Inc., 2002.

[77] Y. Ho, K.-M. Li, and S.-J. Wang, "Leakage monitoring technique in near-threshold systems with a time-based boot-strapped ring oscillator," in *Test Symposium (ATS), 2013 22nd Asian*, Nov 2013, pp. 91–96.

[78] 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 *Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on*, Nov 2002, pp. 721–725.

[79] M. Pons, E. Barajas, D. Mateo, J. Gonzalez, F. Moll, A. Rubio, J. Abella, X. Vera, and A. Gonzalez, "Fast time-to-market with via-configurable transistor array regular fabric: A delay-locked loop design case study," in *Design Technology of Integrated Systems in Nanoscale Era (DTIS), 2011 6th International Conference on*, April 2011, pp. 1–6.

[80] P. Flatresse, B. Giraud, J. Noel, B. Pelloux-Prayer, F. Giner, D. Arora, F. Arnaud, N. Planes, J. Le Coz, O. Thomas, S. Engels, G. Cesana, R. Wilson, and P. Urard, "Ultra-wide body-bias range ldpc decoder in 28nm utbb fdsoi technology," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, Feb 2013, pp. 424–425.

[81] G. Palumbo and D. Pappalardo, "Charge pump circuits: An overview on design strategies and topologies," *Circuits and Systems Magazine, IEEE*, vol. 10, no. 1, pp. 31–45, First 2010.

PART II

APPENDIXES

APPENDIX A

PAPER 1: MONITOR STRATEGIES FOR VARIABILITY REDUCTION CONSIDERING CORRELATION BETWEEN POWER AND TIMING VARIABILITY

Joan Mauricio, Francesc Moll and Josep Altet.

Monitor strategies for variability reduction considering correlation between power and timing variability

J. Mauricio, F. Moll, J. Altet
Universitat Politècnica de Catalunya, Barcelona, Spain
juan.mauricio@upc.edu, francesc.moll@upc.edu, josep.altet@upc.edu

Abstract— **As CMOS technology scales, Process, Voltage and Temperature (PVT) variations have an increasing impact on, performance and power consumption of the electronic devices. Variability causes an undesirable dispersion of performance parameters and a consequent reduction in parametric yield. Monitor and control techniques based on BB and VS can be used to reduce variability. This paper aims to determine which type of sensor provides a better overall variability reduction by taking into account the correlation between different performance magnitudes: static power, dynamic power and delay.**

Keywords: CMOS Digital design; nanometer technologies; leakage; delay; variability; on-chip sensing; critical path; body biasing; voltage scaling

I. INTRODUCTION

The scaling of the CMOS technology has many advantages: more functionality, less area per IC and less cost per unit, among others. On the other hand, this scaling has triggered Process, Voltage and Temperature (PVT) variability issues, producing a decrease in yield.

Variability can be classified in terms of source, scale and nature of variations. The source of variability can be static or dynamic. Static variations are defined at manufacturing time (mainly process variations), and dynamic variations depend on time and environment (supply voltage, temperature, aging). Different variation mechanisms may cause variations at different scale: Within-Die (WID), Die-to-Die (D2D), Wafer-to-Wafer (W2W). Finally, the nature of variations can be random or systematic. Random variability depends on device structure and systematic variability depends on process.

The study of variability has gained prominence in newer CMOS technologies. In new technologies, geometrical variability begins to be dominated by random mechanisms such as Line Edge Roughness (LER), Random Dopant Fluctuations (RDF), among others [2]. Therefore, in the near future, it is expected a wide variability even in nearby transistors.

On-chip sensing is becoming a must in order to detect the effects of PVT variability. On the one hand, built-in delay measurement circuits allow detecting path delay faults that may appear in silicon die due to PVT variability or aging issues. In [3] a novel technique for delay sensing with minimal area and power overhead is proposed. Power sensing is also becoming usual in this new scenario since the scaling of CMOS technology has led to an increase of the power density in a silicon die. This fact combined with PVT variability, can dramatically change the hotspot locations within the chip triggering a thermal runaway [4].

Body Biasing and Voltage Scaling techniques are commonly mentioned in the literature in order to control the effects of PVT variability: In [6] a fully-analog BB calibration method to reduce the mismatches in transistor pairs is proposed. In [7] bidirectional BB is applied to maximize the number of dies accepted in the highest frequency bin. In [8,9] BB and VS techniques are proposed to improve the yield. However, in previous works variability reduction is not considered in all the possible magnitudes (static power, dynamic power and leakage) When variability in one magnitude is reduced, the other magnitudes are affected depending on their variability correlation.

The objective of this paper is to study what sensing strategy (based on delay, dynamic or static power) gives a better overall variability reduction, not only in the magnitude being sensed, but in the other magnitudes as well. We consider a system-level adaptive solution to reduce variability using observable magnitudes (leakage, dynamic power and delay) and adaptive techniques: Body Biasing (BB) and Voltage Scaling (VS). We base our results on Monte Carlo simulations of small and medium sized circuits (8-bit Ripple Carry Adder and 16-Bit Multiplier).

The structure of the paper is as follows. Section II presents several simulations to evaluate power and delay variability as well as their correlation. In Section III the effects of BB and VS on power consumption and delay are evaluated. In Section IV it is analyzed how to best combine BB and VS techniques and which observable magnitude compensates variability more efficiently with respect to nominal values. Finally, section V concludes the paper.

II. VARIABILITY ANALYSIS

The aim of this section is twofold. First, to analyze the sensitivity of delay and power consumption as a function of the parameters susceptible to change due to PVT variations. This analysis will indicate which parameter variation affects the most power consumption and performance. Second, a statistical power and delay analysis is done. This analysis will show the statistical behavior of a given circuit in order to evaluate the impact of PVT variations in terms of power consumption and delay. Correlation between the magnitude's variability can also be computed. Assuming PVT variations are

random as a first approximation, it is possible to simulate the effect of variations by means of Monte Carlo simulations with independent and simultaneous variation of all the parameters.

Simulations were performed using BSIM4 device models for a 65 nm technology. TABLE I. shows the 3-σ variability considered for each parameter considered in the sensitivity analysis [8]:

TABLE I. 3- σ VARIABILITY FOR EACH PARAMETER

Parameter	L	W_P, W_N	T_{OX}	N_{SUB}	V_{DD}	Temp
3-σ var. [%]	±20%	±5%, ±8.5%	±10%	±10%	±10%	80ºC ±50%

A. *Independent Parameter Variations*

The testbench circuit used to perform these simulations is a chain of 15 inverters. For each parameter of the above table, a ±3σ sweep is done keeping the other parameters to its nominal value. To compute static power the chain input is tied to ground, whereas dynamic and delay have been obtained by applying a 1 GHz clock signal to the chain input. The main results of the analysis are in Figure 1 and Figure 2:

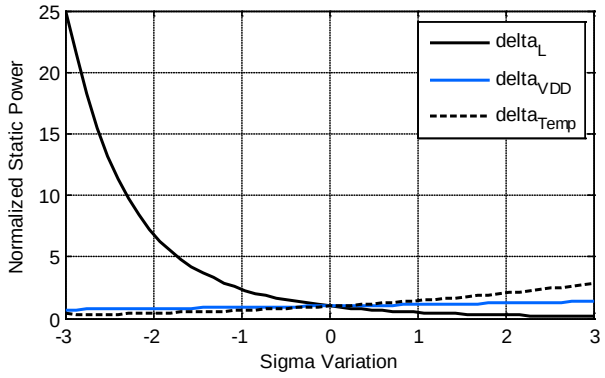

Figure 1: Effects of ±3σ variations on normalized leakage power in a 15 inverters chain

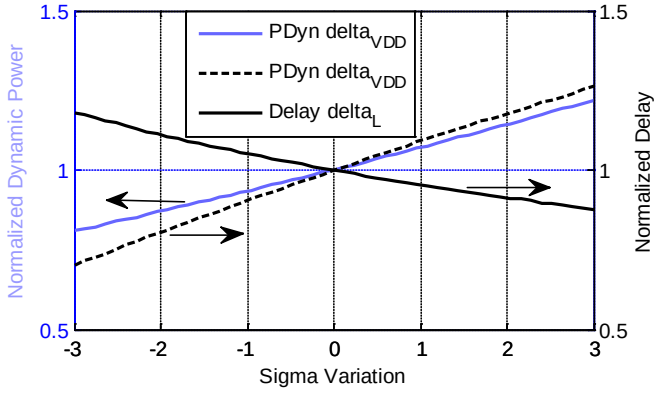

Figure 2: Effects of ±3σ variations on normalized dynamic power and normalized delay in a 15 inverters chain

Figure 1 shows the normalized static power variation (normalized to the nominal case) as a function of the sigma variation of three parameters listed in TABLE I. : L, V_{DD} and T. The other parameters are not included in the figure as their variation induces a change in the normalized static power less than ±5% for a 3 sigma sweep. As it can be seen, gate length variations are the most important source of leakage variations, followed by temperature. Hence, as leakage depends considerably on temperature, environmental factors have an impact on leakage power variability. Although channel width, oxide thickness and channel doping variations do not affect significantly leakage power variability, in future sub-20 nm technologies oxide thickness and channel dose will become critical [2].

Figure 2 shows that dynamic power mainly depends on supply voltage (the other parameters change the normalized dynamic power in less than ±5%). Therefore, dynamic power sensors are suitable to track changes in the supply voltage but are not good process variability indicators. Figure 2 also shows a quasi-linear dependence between channel length, supply voltage and delay. Therefore, delay sensors are suitable to track either process variations or changes in supply voltage.

To sum up, it can be observed that channel length is the major source of process variations, and dynamic variability has a big influence on power and delay variability. As a last remark, PVT variability affects unequally each magnitude, thereby in some cases non-observed magnitudes cannot be well estimated from the observed magnitudes. For example, while channel length variability may produce huge variations in leakage power, dynamic power will not experience significant changes, and therefore, if dynamic power sensors are used, leakage power variability cannot be reduced.

B. *Simultaneous Variation of Device Parameters*

In this case, an 8-Bit Ripple-Carry Adder will be used in order to show a more realistic situation with a larger path with more and different gates, not only single-size inverters.

250 Monte Carlo simulations were performed applying the ±3σ parameter variability shown in TABLE I. . Figures 3 and 4 show the normalized delay, static and dynamic power Cumulative Distribution Function (CDF) of 250 Monte Carlo simulations. In these graphs, the three magnitudes are normalized to the nominal value.

TABLE II. MEAN AND VARIANCE OF STATIC, DYNAMIC AND DELAY

Mean			**Variance**		
Delay	Leakage	Dynamic	Delay	Leakage	Dynamic
1.010	1.946	1.017	1.550e-2	9.071	8.875e-3

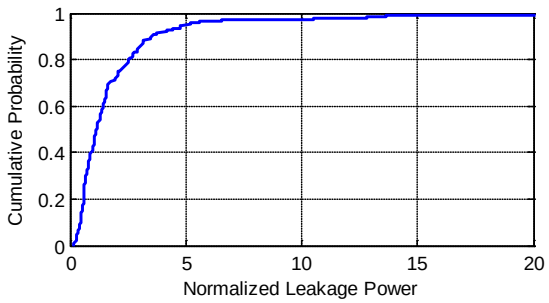

Figure 3: 8-Bit Ripple-Carry Adder static power CDF

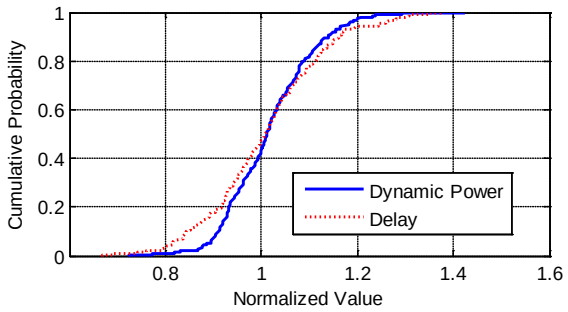

Figure 4: 8-Bit Ripple-Carry Adder dynamic power and delay CDF

Figure 3 shows a big spread in the static power due to PVT variations: around 25% of the circuits present more than 2X nominal leakage. As analyzed in the previous section, this big spread is mainly caused by channel variations, which is the most critical source of leakage power variations.

In Figure 4 showing the distribution of dynamic power and delay, it can be seen that the variations are smaller. In these magnitudes, the distribution is similar to a normal distribution. This can be attributed to the quasi-linear dependence between dynamic power and delay and PVT parameters. PVT variations are defined as normal distributions in the Monte Carlo analysis (one for each parameter), and therefore, a quasi-linear combination of these variations will produce a quasi-normal probability distribution.

As explained in the introduction, the aim of this work is to determine which type of sensor provides the largest overall reduction of variability. In this context, correlations between static power, dynamic power and delay variability play an important role. Control of variations is based on a sensor measuring a given magnitude (e.g. delay). Then, by using control mechanisms, circuits with measured characteristics different from the nominal value are brought to the nominal value and variability is thus reduced for the observed magnitude. A high correlation between magnitudes implies that when using control mechanisms to reduce variability in the magnitude being measured, the variability in the correlated magnitude is also reduced. The higher the correlation between magnitudes is obtained, the higher clock frequency or power consumption variability reduction can be achieved when adaptive techniques are applied. In other words, when no correlation exists between two magnitudes, the application of adaptive techniques based on the observed magnitude will increase the second magnitude variability, rather than bringing it to its nominal value.

TABLE III. shows the correlation coefficient of 250 Monte Carlo simulations:

TABLE III. CORRELATION COEFFICIENTS BETWEEN MAGNITUDES

Correlation	Value
Dynamic Power vs. Static Power	0.49
Delay vs. Static Power	0.62
Delay vs. Dynamic Power	0.70

As shown in TABLE III. static power cannot be accurately predicted from dynamic power measurements, and vice-versa, since their correlation is 49%. On the other hand, delay is better correlated with the other two magnitudes. Therefore, using a delay sensor in order to drive the adjustment can be more beneficial than using power sensors because with delay information also the dissipated power can be adjusted.

III. EFFECT OF ADAPTIVE TECHNIQUES (VS/BB) ON PERFORMANCE PARAMETERS

As mentioned above, Body Biasing (BB) and Voltage Scaling (VS) can be applied to those circuits that do not have a nominal value for the observed magnitude (static power, dynamic power, or delay). The changes in threshold and supply voltage achieved by BB and VS will not only affect the observed value. Hence, variability reduction in the observed magnitude may imply an increase in variability of non-observed magnitudes.

In this section simple analytical expressions and simulation results are presented in order to get an idea of how BB and VS affect the observable magnitudes: static power, dynamic power and delay.

Body bias will cause changes on V_{TH} as given by expression [9]:

$$V_{TH} = V_{FB} + |2\phi_P| + (\lambda_B/C_{OX})\cdot(2qN\varepsilon_S\cdot|2\phi_P|-V_{BS})^{1/2} - \lambda_D\cdot V_{DS} \quad (1)$$

where V_{BS} is the voltage applied to the bulk when ABB is used. When V_{BS} is higher than zero (the effective V_{TH} decreases) in NMOS transistors and lower than V_{DD} for PMOS transistors, the technique is called Forward Body Biasing (FBB), while for V_{BS} is lower than zero (the effective V_{TH} increases) in NMOS transistors and higher than V_{DD} for PMOS transistors, it is called Reverse Body Biasing (RBB).

Leakage power consumption can be expressed as [2]:

$$P_{Leak} = V_{DD}\cdot I_0\cdot(1-\exp(-V_{DS}/V_T))\cdot\exp((V_{GS}-V_{TH})/(n\cdot V_T)) \quad (2)$$

Leakage power depends mainly linearly on supply voltage (V_{DS} also changes when V_{DD} changes but has little influence), and exponentially on threshold voltage. Leakage increases with V_{DD} and FBB, and decreases with RBB.

Circuit dynamic power consumption can be expressed as:

$$P_{Dynamic} = V_{DD}^2\cdot f_{CLK}\cdot\mathrm{sum}(A_{Node}\cdot C_{Node}) \quad (3)$$

where A_{Node} C_{Node} is the capacitance of each node in the circuit weighted by its switching activity. Regarding control variables it is then seen that dynamic power only depends on supply voltage and it is largely independent on threshold voltage. Thus, body biasing does not affect dynamic power consumption.

Transistor delay can be expressed as [10]:

$$\mathrm{Delay} = K\cdot(V_{DD}/(V_{DD}-V_{TH})^{\alpha}), \quad (4)$$

where α is in the range of 1.2 to 1.6 for present technologies, and K depends on the transistor size. This equation clearly shows that delay is reduced either when supply voltage increases, or when FBB is applied.

In order to determine how BB and VS techniques affect power and delay, a 2D parametric analysis was performed in a chain of 15 inverters using BSIM4 device models for 65 nm technology. In this analysis both VS and BB techniques are applied. Figure 5 to Figure 7 show static power, dynamic power and delay with respect to VS and BB respectively.

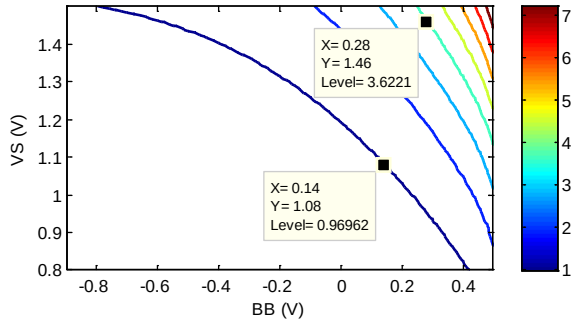

Figure 5: Normalized static power consumption vs. BB & VS

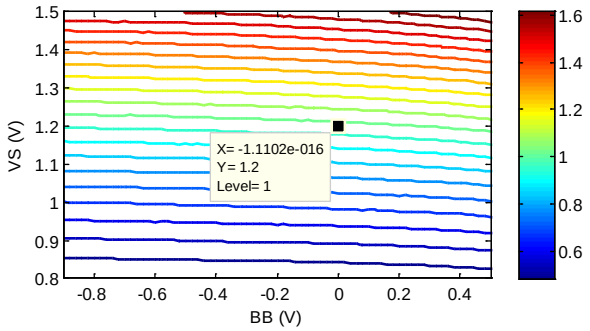

Figure 6: Normalized dynamic power consumption vs. BB & VS

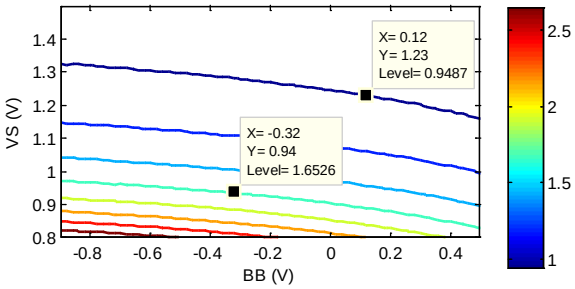

Figure 7: Normalized delay vs. BB & VS

From the level curves in Figure 5 to Figure 7 it can be observed that, in agreement with leakage power, dynamic power and delay analytical expressions, static power is increased either when V_{DD}, V_{BS} (or both) increases, while delay is reduced either when V_{DD}, V_{BS} (or both) increases.

As shown in Figure 5 the best way to reduce leakage power and to not increase delay significantly is to apply RBB. In the cases when leakage power is much larger than the nominal value, RBB must be combined with negative VS at the expense of an increase in delay.

On the other hand, if leakage is smaller than the nominal value, delay will be probably higher than expected. Then, two approaches are possible. One is to apply FBB so that leakage power will increase significantly while delay will be moderately reduced. Dynamic power will not be affected. The second alternative approach is to increase V_{DD}, so that leakage power will increase a little, while delay will be reduced considerately. Dynamic power will also be increased considerately. Hence, depending on the priorities of the variability to be reduced, FBB, an increase of V_{DD}, or both should be applied.

As can be observed in Figure 6, the most effective way to adjust dynamic power is to modify V_{DD} since BB doesn't have relevant impact on dynamic power consumption.

In Figure 7 it can be observed that circuits with a delay slightly below the nominal value will benefit of applying RBB as this strategy will reduce leakage as well. If a larger adjustment is necessary, negative VS can also be applied.

For circuits having a delay larger than nominal, either positive VS or a combination of FBB with positive voltage scaling should be applied.

IV. VARIABILITY REDUCTION WITH DIFFERENT SENSORS

A sensor is needed in order to apply adaptive techniques. Once it is seen the statistical behavior of power and delay in the presence of PVT variability, and adaptive techniques have been introduced, four different approaches are evaluated in order to determine which kind of sensor gives a better performance, in terms of variability reduction:

1. Leakage power sensor.
2. Dynamic power sensor.
3. Delay sensor.
4. Dual sensor: Dynamic power + Delay.

The algorithm flow chart is shown in Figure 8. Maximum and minimum BB and VS values are shown in TABLE IV. :

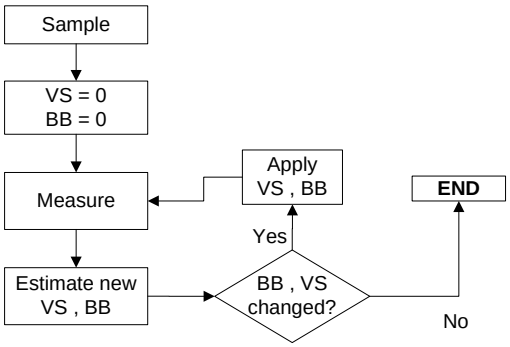

Figure 8: Variability reduction algorithm

TABLE IV. MINIMUM AND MAXIMUM BB AND VS VOLTAGES

Parameter	Nominal	Min	Max
BB	0V	-0.9V	0.5V
VS	1.2V	0.8V	1.5V

The algorithm shown in Figure 8 describes the adjustment process which takes place in a circuit in the presence of PVT variations. First, the observable magnitude is measured. Then, body bias (BB) and supply voltage (VS) values are estimated. BB and VS are those values which bring measured value to nominal value. If computed BB and VS values don't differ from the previous values, circuit is already adapted and nothing has to be done. Otherwise, estimated voltages are applied and the observable magnitude of the circuit is measured again, and a new estimation is done. This iterative process ends when the previous and the new computed BB and VS values are equal within a given resolution. For this work, body bias and supply voltage resolutions are ±0.01V and 0.005V respectively.

The estimation process consists on computing the magnitude gain to be applied at the circuit to bring them to its nominal behavior. This gain is defined as observed magnitude divided by nominal value. Estimated BB and VS values are obtained from the level curves shown in Figure 5 to Figure 7. The algorithm finds the new BB and VS values which best fit with the desired magnitude gain. Despite the level curves were obtained from a nominal 15 inverters chain, the estimation process was also successful when larger circuits such as ISCAS C6288 were simulated. This is possible since power and delay level curves do not change substantially between different circuits, and therefore, these level curves are enough to estimate the appropriate BB and VS values.

In case of dual sensors the process is similar: the algorithm tries to find those BB and VS values which delay and dynamic power gains best fit with the desired ones. If neither BB nor VS values can achieve these restrictions, the algorithm finds those values which delay and dynamic power gains are the nearest to the desired ones.

250 Monte Carlo simulations are performed applying the ±3σ parameter variability shown in TABLE I. , using an 8-Bit Ripple-Carry Adder.

Figure 9 to Figure 11 show the Cumulative Distribution Function (CDF) of 250 Monte Carlo simulations when no adaptive techniques are used (bold line), and when static power, dynamic power, delay or dual sensor is used to reduce variability. CDF indicates the percentage of cases which accomplishes that power or delay is equal or less than a certain value.

Although leakage sensing achieves excellent results in terms of leakage power, this reduction of leakage variability implies a worsening of statistical performance in terms of delay, which implies parametric yield loss and, what is more, a substantial increase of mean dynamic power variability. This can be attributed to the low correlation between leakage power and delay and, particularly, between leakage and dynamic power. Leakage is mostly a problem during inactivity periods. In these periods a better approach would be to shut off the block avoiding leakage altogether. Therefore, the efforts should be concentrated on those periods when the system is running.

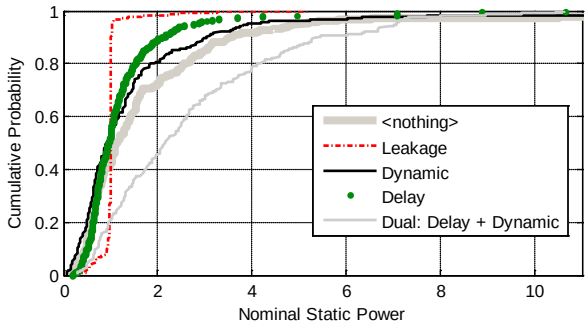

Figure 9: Nominal static power CDF in an 8-Bit Ripple-Carry Adder applying variability reduction

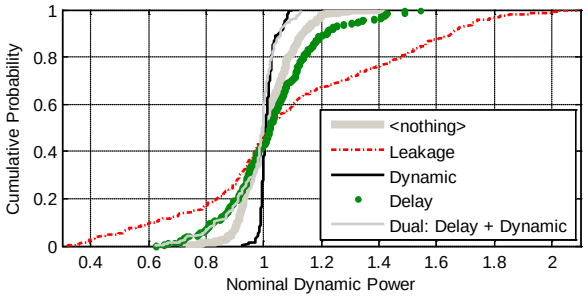

Figure 10: Nominal dynamic power CDF in an 8-Bit Ripple-Carry Adder applying variability reduction

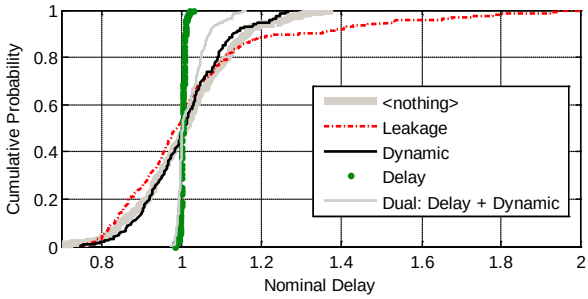

Figure 11: Nominal delay CDF in an 8-Bit Ripple-Carry Adder applying variability reduction

Dynamic power sensor does not contribute to reduce PVT effects substantially. Although dynamic power variability is reduced, delay and static power do not have remarkable improvements.

As expected, Delay sensors provide an excellent delay variability reduction without worsening dynamic power consumption variability. Leakage power variability is also substantially improved.

Dual sensor strategy achieves the best trade-off between dynamic power and delay. In the 90% of cases, dynamic power consumption and delay CDF is equal or less than +5% and +7% of nominal dynamic power and nominal delay respectively. On the other hand, leakage power consumption experiences an increase in terms of mean and variance. However, leakage power consumption is not an issue when no power restrictions exist. What is more, if some power management strategy is also implemented leakage power consumption can be reduced during the idle periods [5].

V. CONCLUSIONS

Channel length, temperature and supply voltage are the most important source of variations. Nevertheless, these parameter variations affect differently static power, dynamic power and delay. For this reason, when two by two correlation coefficients are computed in Monte Carlo simulations, ranges of results are between 49% and 70%. The results depicted in this paper are very similar for small circuits such as a 15 inverters chain or an 8-Bit Ripple-Carry Adder, or more complex circuits such as a 16-Bit Multiplier (ISCAS C6288). It is reasonable to expect that the conclusions are also applicable to large SoC circuits.

A novel study to reduce variability is proposed. This study evaluates the impact of variability reduction in different magnitudes (static and dynamic power and delay) when adaptive techniques are used to minimize each of them. It has been shown that when using one type of sensor to control the sensed magnitude, the magnitudes not being directly sensed are also affected when adaptive techniques applied (BB and VS) and sometimes the resulting yield is worsened. This is due to correlation between magnitudes. In this work this correlation is taken into account and quantified.

Delay is the magnitude which has better correlation coefficients, 62% with leakage power and 70% with dynamic power. This fact explains the good results obtained when delay is sensed and used as feedback to reduce variability. Higher correlations between delay and power mean that PVT variations affects similarly and, therefore, when delay variability is reduced by means of BB or VS, static and dynamic power variability not increase and, in some cases, may even be reduced .

Dual sensors also provide a good trade-off between delay and dynamic power variability, and may be a good choice for those systems which its leakage power consumption is not an issue. Otherwise, an additional power management technique such as power gating will be mandatory.

Finally, delay sensors are proposed in order to implement variability-tolerant smart circuits, rather than the other three approaches. However, delay sensors may not be always the best choice for reducing variability. In some specific low power applications without severe timing and dynamic power restrictions, leakage power sensors may be a better choice, for example. The aim of this work is to show that variability can be afforded with different solutions. However, delay sensors give better results in general terms, with less extra hardware and lower area overhead than other approaches.

TABLE V. shows statistical results for circuits of different sizes when delay sensors are used to reduce variability (right columns), and when no adaptive technique is applied (left columns).

TABLE V. VARIABILITY STATISTICS WHEN DELAY SENSORS ARE USED

	Normalized Mean		**Normalized Variance**		**3-σ / μ [%]**	
Circuit	15 inverters chain (30 gates)					
Delay	1.00	1.00	1.33e-2	3.68e-5	34.6	1.82
Dynamic Power	1.01	1.02	5.43e-3	1.97e-2	22.1	42.1
Static Power	2.14	1.49	7.49	1.40	821	355
Circuit	8-Bit Ripple Carry Adder (224 gates)					
Delay	1.01	1.00	1.55e-2	3.31e-5	37.4	1.69
Dynamic Power	1.02	1.02	8.87e-3	2.52e-2	28.3	47.7
Static Power	1.95	1.39	9.07	1.40	904	354
Circuit	16-Bit Multiplier (C6288) (2406 gates)					
Delay	0.98	1.00	1.28e-2	3.87e-5	34.0	1.87
Dynamic Power	1.11	1.09	2.6e-2	3.15e-2	48.4	49.0
Static Power	1.89	1.08	6.06	0.35	738	165

ACKNOWLEDGMENT

This work was partly supported by project MODERN funded by the Spanish MICINN through ``Fondo Especial para la Dinamización de la Economía y el Empleo -- Plan E" (contract PLE2009-0024) and the ENIAC JU (contract ENIAC-120003) and TERASYSTEMS TEC2008-01856.

REFERENCES

[1] Devgan, A.; Nassif, S.; , "Power variability and its impact on design," *VLSI Design, 2005. 18th International Conference on* , vol., no., pp. 679- 682, 3-7 Jan. 2005.

[2] Asenov, A.; , "Alternatives and Characteristics of Technologies sub-20nm, " *Course Series in Advances on Reliable Digital Design*, UPC 6-7 July 2010. Available at: http://doctorat.eel.upc.edu/research/courses-and-seminars.

[3] Datta, R.; Sebastine, A.; Abraham, J.A.; , "Delay fault testing and silicon debug using scan chains," *Test Symposium, 2004. ETS 2004. Proceedings. Ninth IEEE European* , vol., pp. 46- 51, 23-26 May 2004.

[4] Jaffari, J.; Anis, M.; , "Statistical Thermal Profile Considering Process Variations: Analysis and Applications," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* , vol.27, no.6, pp.1027-1040, June 2008.

[5] Martin, S.M.; Flautner, K.; Mudge, T.; Blaauw, D.; , "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads," *Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on* , vol., no., pp. 721- 725, 10-14 Nov. 2002.

[6] Bacinschi, P.B.; Murgan, T.; Koch, K.; Glesner, M.; , "An Analog On-Chip Adaptive Body Bias Calibration for Reducing Mismatches in Transistor Pairs," *Design, Automation and Test in Europe, 2008. DATE '08* , vol., no., pp.698-703, 10-14 March 2008.

[7] Tschanz, J.W.; Kao, J.T.; Narendra, S.G.; Nair, R.; Antoniadis, D.A.; Chandrakasan, A.P.; De, V.; , "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *Solid-State Circuits, IEEE Journal of* , vol.37, no.11, pp. 1396- 1402, Nov 2002.

[8] Chen, T.; Naffziger, S.; , "Comparison of adaptive body bias (ABB) and adaptive supply voltage (ASV) for improving delay and leakage under the presence of process variation," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* , vol.11, no.5, pp. 888- 899, Oct. 2003.

[9] Srivastava, A.; Bai, R.; Blaauw, D.; Sylvester, D.; , "Modeling and analysis of leakage power considering within-die process variations," Low Power Electronics and Design, 2002. ISLPED '02. Proceedings of the 2002 International Symposium on , vol., no., pp. 64- 67, 2002.

[10] Figueras, J.; Macii, E.; , "Leakage Power Modeling and Control Strategies" *Clean Workshop*, Lisbon 8 September 2008. Available at: http://tec.upc.es/sed/Leakage%20power%20basics.ppt.

APPENDIX B

PAPER 2: MEASUREMENTS OF PROCESS VARIABILITY IN 40 NM REGULAR AND NON-REGULAR LAYOUTS

Joan Mauricio, Francesc Moll and Sergio Gómez.

Measurements of Process Variability in 40 nm Regular and Non-Regular Layouts

Joan Mauricio, Francesc Moll *Member, IEEE.* and Sergio Gómez

***Abstract*— As technology scales down, IC design is becoming more difficult due to the increase in process variations, which translates into a dispersion of circuit parameter values thus degrading manufacturing yield. Regular layouts are recommended to reduce variability with the cost of area overhead with respect to conventional layouts. The aim of this paper is to measure the impact of variability in two implementations of the same circuit in a commercial 40 nm technology: a regular layout style and a compact, non-regular layout. Experimental results show a 60% reduction in variability with a cost of 60% area overhead.**

***Index Terms*— Variability, Lithography distortion.**

I. Introduction

The progress of the semiconductor industry and its impact on society has been mainly based on the continuous scaling of technology. Scaling has enabled an increase in device density that implies a larger functionality of individual chips and also, traditionally, an improvement in device characteristics. However, as the characteristic device dimensions enter the nanometer scale, the device improvement is becoming more difficult to achieve due to manufacturing challenges at such small dimensions [1].

One major difficulty of the technology scaling in current and upcoming technologies is the increase in variability [2]. Usually process variability refers to variations across the wafer, from wafer to wafer and from lot to lot due to imperfection of process control and equipment nonuniformity. Nowadays, variability becomes increasingly important due to a combination of factors. First, since nominal dimensions are extremely small, physical parameter fluctuations have a large relative impact. This source of variability is known as random or statistical and it is local in nature, giving rise to device-to-device variations even in devices with identical layout. Examples include Random Dopant Fluctuations, Line Edge Roughness, among others [3], [4]. Second, fluctuations or non-homogeneities of the manufacturing process also contribute to the overall indeterminacy of each device characteristics. This second source of variation is known as systematic and it can produce both local (device-to-device) and global (die-to-die) variations. Examples of systematic variability sources include lithography distortion, flare, non-uniform layer thickness due to CMP and well proximity effect.

The authors are with the High Performance IC Design Group, Dept. of Electronic Engineering, Universitat Politècnica de Catalunya.

The research leading to these results has received funding from the European Community Seventh Framework Programme (FP7/2007-2013) under grant agreement number 248538 (MODERN project) and the Spanish Ministry of Economy (MINECO) and FEDER funds through the project TEC2008-01856 (Terasystems).

Variability represents a technical, but especially an economical problem. Device variability translates into a dispersion of circuit specifications so that a certain number of circuits do not accomplish the desired specifications and thus the yield decreases. As a consequence, variability mitigation techniques are applied at different levels. While random variability cannot be completely eliminated, new device architectures tend to reduce its effects for example by reducing or eliminating channel doping, as seen in FD-SOI [5] or FinFETs [6]. With respect to systematic variability, regular layout design styles [7], [8] are used to minimize the number of patterns in each layer and to simplify the patterns by using only rectangular shapes and thus reducing the systematic variability associated to lithography distortion. These techniques aim to reduce local variations, while global variations caused by other systematic or general process variability can be addressed by electrical tuning (voltage scaling, body bias, etc.) [9].

While regular layout techniques are usually recommended to reduce variability, they also represent an area overhead with respect to conventional layout. On the other hand, the benefits in terms of variability of regular layouts have not been adequately quantified and for this reason it is difficult to assess whether regular layouts represent an overall advantage with respect to conventional layouts or not.

The objective of this paper is to measure the impact of variability in two different implementations in a commercial bulk CMOS 40 nm technology of a Voltage Controlled Delay Line (VCDL) circuit: one in a highly regular layout style called VCTA [10], and the other using a compact, non-regular layout design style. The VCDL circuit is chosen as a test vehicle, so that delay is measured reflecting the different sources of variability present in the circuit. The variability measurements can contribute to better assess the advantages and disadvantages of a regular layout design style against a conventional one.

The structure of the article is as follows. Section II briefly outlines the regular layout used for the comparison, called VCTA. Section III describes the chip implementation and the different design styles used to assess variability effects. In Section IV the types of measurements to evaluate variability are detailed. Section V shows the results of Global Variations measurements whereas Section VI presents the results of Local Variations in VCDL delay stages. Finally, Section VII concludes the paper.

II. VCTA Regular Layout

The VCTA proposal, described in [10], is a very regular structure that maximizes layout regularity by setting up regular

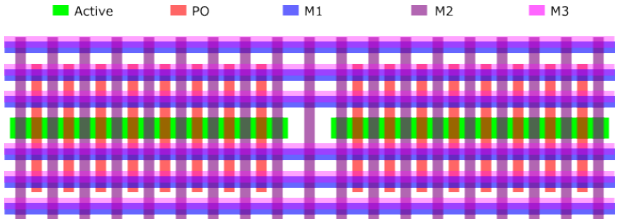

Fig. 1. Basic VCTA cell.

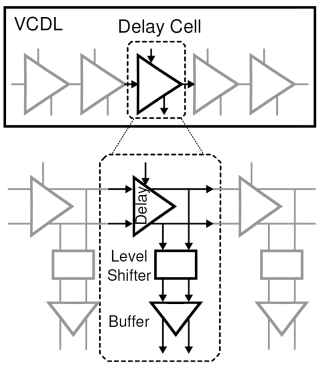

Fig. 2. Schematic of the VCDL.

interconnects and enforces all transistors to have the same dimensions. It is based on a basic cell that contains 6 PMOS and 6 NMOS transistors, and an interconnection grid from M1 to M3 levels (see Figure 1). The connectivity of the transistors is done by proper use of contacts and vias.

The advantages of VCTA are twofold. On the one hand, the constant pitch between metals as well as poly layers ensures the minimum lithography variability due to neighbor proximity effects. On the other hand, the geometry of the VCTA cell eases the layout automation and its scalability.

The main disadvantage is the area penalty due to the use of dummy transistors. There are also performance losses associated to the large parasitic coupling capacitance due to dense and long parallel interconnection lines.

III. Chip Structure

A. VCDL Design

A VCDL is a chain of M delay elements (buffers) such that its total delay can be tuned by adjusting its supply voltage (Control Voltage, VCNT) to obtain a fixed delay between the ends of M stages. Assuming that the VCDL area is small enough so that global variations do not affect one single chain, local variations will be the main source of variability from chain to chain. Hence, each of the VCDL stages has its own independent delay. A previous work [11] used this circuit simulated in a 90 nm technology to evaluate variability.

In this work, a differential signal VCDL has been designed and fabricated in bulk 40 nm technology, as shown in Figure 2. For each delay element there is a level shifter that restores the differential output voltages from VCNT to VDD. In this way the differential output signals of each stage have signals of amplitude equal to VDD. Finally, the buffer reduces rise and fall time after level conversion.

Both VCNT and VDD supply distribution have dedicated power pads.

B. Design Styles

In order to evaluate the differences between regular and non-regular layouts, two layout implementations were created. On the one hand, we designed a VCTA implementation of a VCDL. On the other hand, we designed a non-regular full custom (FC) design, which does not take into account lithography imperfections. The transistor channel area is the same for VCTA and FC styles.

Furthermore, the VCDL was implemented in two sizes for each design style, since according to Pelgrom's law random variability decreases with $\sqrt{WL}$. For each design style, VCDLs are implemented in two versions: one with transistors having the minimum channel width (FC x1, VCTA x1) and another one with transistor width of four times this minimum channel width (FC x4, VCTA x4). In this way, it will be possible to determine the dependence of local variations as a function of transistor size and thus determine the relative influence of layout dependent variations against random variations.

C. VCDL Multiplexing and Chip Floorplaning

In order to have a reasonable amount of statistics and taking into account that the number of chip samples was limited (22 samples), every chip contains several instances of each VCDL version. The total chip area (and therefore the total number of delay stages) was constrained to 1 square millimeter. Hence, a trade-off existed between the number of VCDLs and the number of stages per VCDL. The longer the VCDL, the larger are the local variations that can be measured, at the expense of the statistical population. The number of stages of each VCDL was decided at design time to be 64 stages and the number of VCDL blocks to be 8, consecutively arranged in a row. This gives a total population of 8x22 (176) VCDLs for each layout style and transistor size.

The four design styles (FC x1, FC x4, VCTA x1 and VCTA x4) have the same block structure, shown in Figure 3. The 8 blocks are connected in cascade, so that a square input signal propagates along the 8 VCDL64 blocks consecutively. Two arbitrary outputs from any VCDL64 block can be selected by setting digital select inputs for MUX64 (O[5:0] and D[5:0]). Finally, a second level of 3-bit multiplexors allows selecting these two outputs of any VCDL64 block.

Figure 4 shows the floorplan location for each of the 8 blocks. Labels VT10 through VT17 correspond to VCTA x1 VCDL64 blocks. Labels VT40 through VT47 correspond to VCTA x4 VCDL64 blocks. Similarly, FC1n and FC4n correspond to FC x1 and FC x4 layout styles.

The use of multiplexors is forced by the large quantity of VCDL64 outputs to monitor (64x8 outputs for each design style, i.e. 2048 outputs altogether) and the limited number of pads. An unwanted side effect of the use of multiplexors is that inevitably they add uncertainties to the measurements,

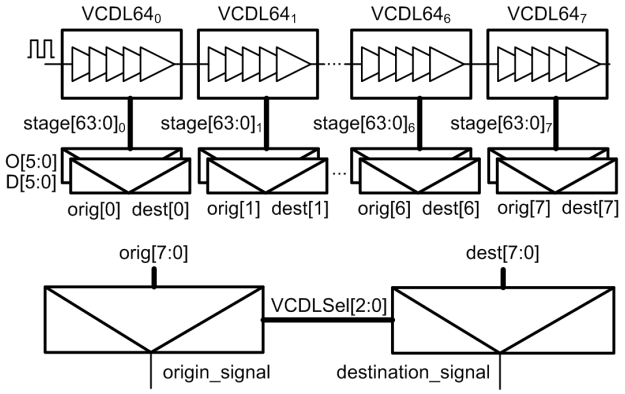

Fig. 3. VCDL multiplexing scheme.

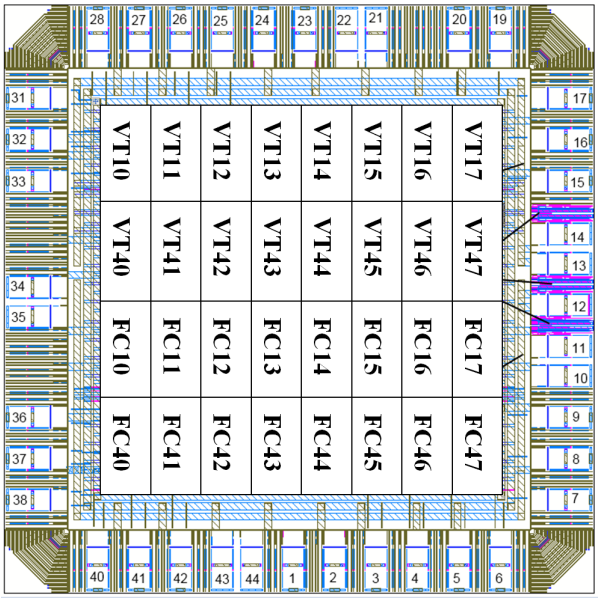

Fig. 4. Chip floorplanning. In the labels the first two characters indicate whether the block is full custom or VCTA; the next one indicates the layout size and the last one indicates the instance number.

since origin and destination signals will experience different delays due to PVT variations. In order to minimize variability in the signal selection, identical instances of VCTA-based multiplexors were used to select VCDL64 signals in all the design styles. In this way the addition of different systematic variations to each implementation is avoided. Therefore, it can be assumed that the observed differences between implementations are principally caused by VCDLs.

Regarding the area penalty, VCTA implementations of the VCDL64 have an area overhead of 60% in comparison with their full custom counterparts, as reflected in Table I. Note that this overhead depends on the utilization of transistors, and this is circuit dependent. In this particular case, the transistor utilization is 70%.

The mean delay is also shown in Table I, normalized to the VCTA x1 delay. As expected, FC style is significantly faster due to smaller wiring parasitics. It is noteworthy that the VCTA x4 is even slower than the VCTA x1 style. This is

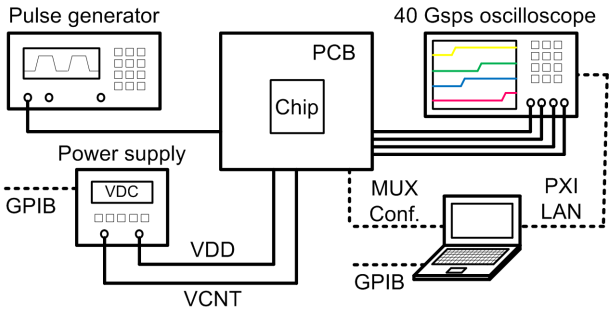

Fig. 5. Testbench setup diagram.

because in order to design the VCTA x4 layout, the VCTA x1 design was replicated 4 times connecting the transistors in parallel, so that the wiring load overcompensates the higher drive capability. It is possible that a more clever design could avoid this effect. In any case, the purpose was to evaluate the effect of channel area on variability and not to optimize the delay, which is only presented here for completeness.

TABLE I
SIZE AND RELATIVE DELAY OF EACH VCDL64 STYLE.

Layout Style	VCTA x1	FC x1	VCTA x4	FC x4
Area (μm^2)	80×20	80×12.5	80×80	80×50
Delay (a.u.)	1	0.68	1.37	0.43

IV. TYPES OF MEASUREMENTS

In order to measure the chip samples a PCB was designed. The diagram of the testbench setup is shown in Figure 5.

A pulse generator provides the square input signal to the chip. Origin and destination stages of each design style are connected to a 40 GS/s oscilloscope. A PC controls the measurement procedure: it configures the DC power supply, generates the multiplexor selection signals, and finally gathers the oscilloscope measurements.

A. Global Variations Measurement Procedure

The first kind of measurements aim to measure the differences in the complete VCDL64 chain delay when working at nominal conditions (1 V). The observed differences reflect die-to-die and on-chip variations, both random and systematic in nature. The specific on-chip variations across chip are also measured by binning the obtained delay by their position in the chip.

In order to measure the effects of process variability, the Control Voltage VCNT and the value of VDD are set to 1 V, 64:1 reference multiplexors (origin) are configured to select the output of stage 0 of a VCDL64 while the other multiplexors (destination) point to the stage 63 of the same block. Once the multiplexors are configured, the oscilloscope measures the difference between the rising edge of the first and the last VCDL64 stage. Finally, the measured value is sent to the PC. This procedure is repeated for each of the 8 VCDL64 blocks by changing the selection bits of the 8:1 multiplexors.

B. Local Variations Measurement Procedure

In this procedure the objective is to determine the difference in delay between individual stages of a given VCDL64 chain when the chain delay is fixed to 10 ns. The differences will be due to local variations (both statistical and systematic associated to regularity), since global variations are eliminated by fixing the total delay.

The measurement procedure of each VCDL64 is done in three steps. Firstly, the value of VDD is set to 1.1 V and the offset delay between multiplexors caused by external sources, such as multiplexor delay variability and cable lengths, is measured by setting the output of both 64:1 multiplexors (origin and destination) to 0. The oscilloscope is configured to measure the difference between rising edges of both multiplexors that will be subtracted in subsequent measurements. Secondly, global process variability effects are uncoupled by adjusting the Control Voltage VCNT so that all the statistical samples have a reference delay of 10 ns between stage 0 and 63 of each VCDL64. This procedure is done by setting the origin of 64:1 multiplexor to 0 and destination multiplexor to 63. Delay is measured after adjusting the output voltage of the VCNT Power Supply. This adjustment process is repeated until the measured delay between stage 0 and 63 is within ± 50 ps of the target delay. Finally, once the offset between multiplexors is compensated and the delay between the first and the last stage of the VCDL64 is set to 10 ns, the origin multiplexor is set to stage 0 and the destination multiplexor is swept from stage 1 through 63 and delay between multiplexors is measured and sent to the PC. This procedure is repeated for each of the 8 VCDL64 blocks.

V. Measurements of Global Variations

As previously explained in Sec. IV-A, the effects of process variability are measured by setting the value of VCNT to 1 V and measuring the time that the signal takes to cross each VCDL64.

A. Overall Variability

In this subsection, we represent a delay histogram of 176 VCDL64 blocks (8 per chip) of each design style in the 22 measured chips. Delays are represented regardless of the physical location of each VCDL64 block. This measurement shows the robustness of each design style with respect to global variations. According to Figure 6, it can be observed that full custom designs experience a wider process variability spreading in comparison with VCTA designs. Additionally, variability is significantly reduced in full custom designs by increasing transistor size, whereas variability only slightly decreases when VCTA layout size increases.

TABLE II
Coefficient of Variation of the Overall Variability.

Layout Style	VCTA x1	VCTA x4	FC x1	FC x4
$3\sigma/\mu$ (%)	8.93	8.46	22.92	15.18

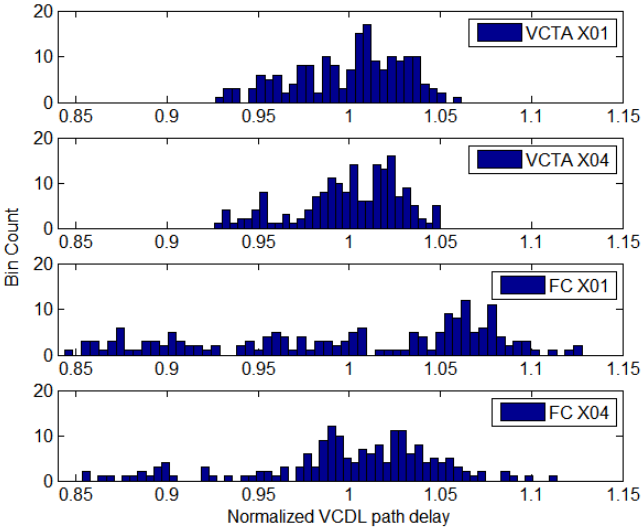

Fig. 6. Histogram of normalized VCDL64 path delay for each design style.

B. Within-Die and Die-to-Die Variability

In this subsection it is shown whether any correlation exists between the physical location of the VCDL64 blocks inside the chip and measured delay, i.e. the evidence of Within-Die (WID) Variations.

In Figure 7 WID Variability is computed by averaging the delays of all the chip samples according to their physical location inside the chip. In other words, the results are the average delays of each block depicted in the chip floorplanning shown in Figure 4. It can be seen that those VCDL64 blocks closer to the die boundary experience a smaller delay (up to 13% with respect to the mean for the FC x1 style) in comparison with those located in the middle. Observing VCDL64 normalized path delay for each design style, we can conclude that VCTA designs suffer less the effects of WID variations than FC.

TABLE III
Ranges of the D2D coefficient of variation according to their location.

Layout Style	VCTA x1	VCTA x4	FC x1	FC x4
$3\sigma/\mu$ (%)	3.66 – 6.05	2.95 – 4.24	4.64 – 6.76	6.48 – 7.47

The spatial correlation of variability is evident in Table III which shows the ranges of the coefficient of variation when the statistical population of the VCDL64 blocks is sorted according to its physical location inside the chip, then effectively giving Die-to-Die (D2D) variability. It is clear that D2D variability is significantly lower than the overall variability shown in Table II. Also, the difference between VCTA and FC styles is not as evident as in the overall results presented in the previous section, especially taking into account that the statistical population used to compute the coefficient of variation is only 22 samples.

VI. Measurements of Local Variations

In this section the objective is to study variations occurring at cell level. As previously explained in Section IV-B, end-to-

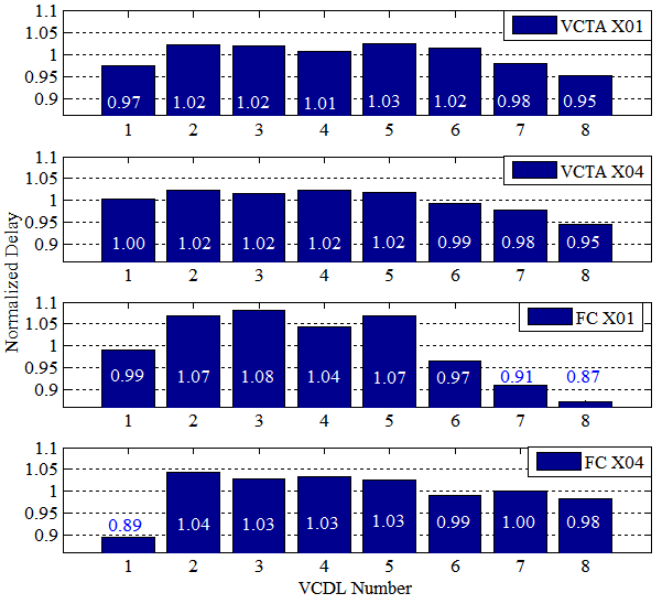

Fig. 7. Normalized average path delay for each of the 8 VCDL64 in each layout design.

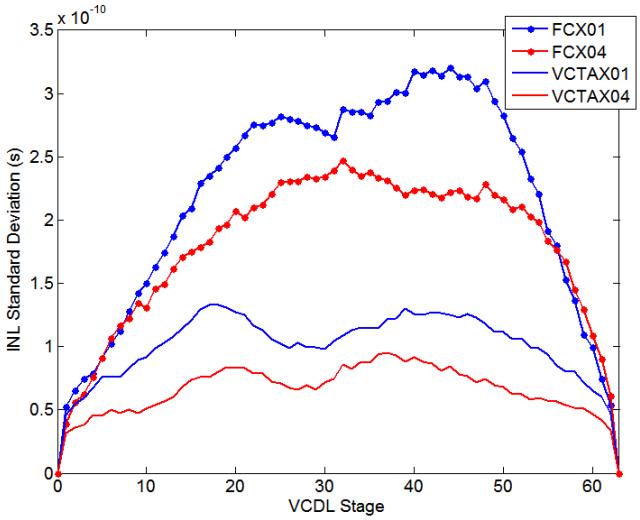

Fig. 8. σ_{INL} for each VCDL64 stage.

end VCDL64 delay is adjusted to the same reference value of 10 ns for all VCDL types, so that any difference between the VCDL64 samples due to global variability is decoupled from the measurements.

We take 64 measurements, one for each stage, in each VCDL64 block. In the absence of variations each VCDL64 is expected to have a delay of $T/63$, where $T = 10$ ns is the end-to-end VCDL64 delay, since the contribution of each stage is expected to be identical. We compute the Integral Non Linearity (INL) of each VCDL64, i.e., the difference between the measured cumulated delay and the expected cumulated delay of each stage:

$$\text{INL}(k) = \text{Delay}(k) - k \cdot 10 \text{ ns}/63 \quad (1)$$

In the absence of variability the 64 INL points of the function would be zero, but the presence of Process, Voltage and Temperature (PVT) variations as well as the oscilloscope resolution will produce fluctuations in this INL. In order to determine the INL we did some adjustments. On the one hand, as explained above, we decoupled variations by adjusting the end-to-end delay of the VCDL64. Additionally, we took all the measurements under similar environmental conditions. Other effects such as multiplexor variability, supply voltage noise and oscilloscope resolution are decoupled a posteriori, once statistical results are computed.

Once the statistical population of 176 VCDL64 blocks is measured, the INL function is computed for each VCDL64 sample and then point-to-point statistical deviation is obtained. Therefore, the resulting vector is the statistical deviation for each VCDL64 stage.

$$\sigma_{\text{INL}}(k) = \sqrt{Var\left[\text{INL}(k)\right]} \quad (2)$$

Figure 8 shows the statistical deviation of the INL for each design style. Transistor local variability (either caused by statistical variability or layout-related effects) produces fluctuations in each VCDL64 stage. Since the initial delay (0 ns) and the end delay (10 ns) are predefined, the maximum delay uncertainty resides in the middle stages. Therefore, we expect semi oval σ_{INL} functions where the minimums are on stages 0 and 63 and the maximum on the stage 32 of the VCDL64. However, the non-uniformity of process variations along the VCDL64 (which strongly depend on its physical location as observed in Figure 7) produces changes on σ_{INL} measurements. Other effects such as multiplexor variability, supply voltage noise and oscilloscope resolution also distorts the measured σ_{INL}.

The next step consists in decoupling local effects from other process variations and the uncertainties introduced by the measurement equipment, and thus estimate the local variations impact in regular and non-regular layouts. The INL added by the power supply and the oscilloscope can be seen as a floor noise since the power supply noise and the oscilloscope resolution accuracy do not vary regardless the VCDL64 stage number. We measured the statistical deviation introduced by the measurement equipment and we determined that the introduced floor noise was 25 ps. Finally, we fit the resulting measurements (the plots from Figure 8 minus the floor noise) by means of the minimum squares method with the following function:

$$\sigma_{\text{INL,FIT}}(x) = A_K \cdot \sqrt{x \cdot (63 - x)} \quad (3)$$

where A_K is the amplitude value which ensures the minimum mean square error between the measurements and the obtained semi oval functions. The correlation coefficient between the measured σ_{INL} (after subtracting the floor noise) and $\sigma_{\text{INL,FIT}}$ are in the range between 91.7% and 94.6%. In Figure 9 the curve fitting results can be observed and Table IV shows the maximum deviation ($\sigma_{\text{INL,FIT}}$) that each design style suffers due to local variations.

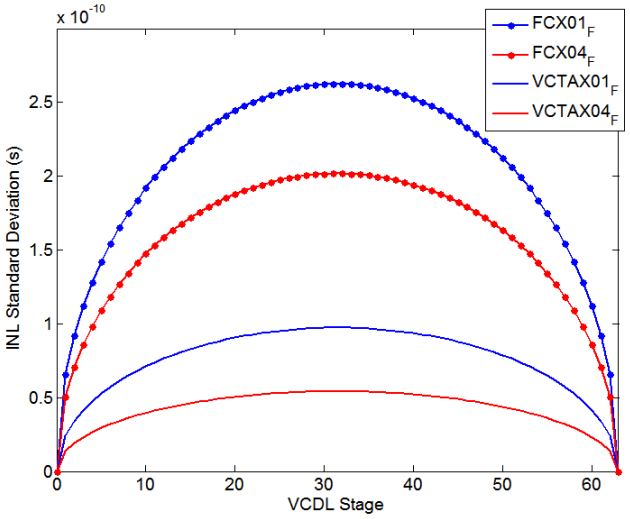

Fig. 9. Estimation of σ_{INL} due to local variations.

Observing the INL in VCTA layouts we can conclude that random variability is the main cause of local variations in regular layouts, since increasing 4 times the layout size, variability improves in a factor of of 1.8 close to factor 2, the expected improvement in random variability according Pelgrom's Law. On the other hand, in full custom layouts local variations are dominated by systematic effects, and consequently, variability only improves a 30% when transistor width is increased by a factor 4. Nevertheless, the improvement in absolute value is similar to VCTA.

In conclusion, regular layouts not only suffer less from local variations than their non-regular equivalents but their variability also scales better.

TABLE IV

MAXIMUM VALUE OF $\sigma_{\text{INL,FIT}}$.

VCTA x1	VCTA x4	FC x1	FC x4
97 ps	54 ps	263 ps	202 ps

VII. CONCLUSIONS

In this paper we presented the measurements of the impact of process variability in two different implementations of a VCDL circuit in a commercial bulk CMOS 40 nm technology. We showed that regular layouts, implemented using the VCTA proposal, exhibit lower variability effects than the non-regular proposal, based on full custom design style.

From the global variability measurements here presented we can observe that the most important part of variability is due to the physical location (within-die variations). These WID variations could be attributed to boundary effects more evident in the FC style, or differences in the power distribution between styles for the design of this particular test chip. Die-to-die variability results show small differences between VCTA and FC designs and therefore we conclude that layout style is not determinant in die-to-die process variability.

From the local variability analysis we can conclude by experimental evidence that the main contribution of local variations in VCTA designs is statistical variability, whereas the main contribution in full custom designs is systematic variability associated to layout. As a consequence, full custom designs not only have bigger local variability than VCTA, but also scale worse with channel width.

The obtained results strongly depend on the technology we used; the more mature a technology is, the less variability it suffers. Given that 40 nm is a mature technology (available since 2008), we can conclude that the differences in variability between regular and non-regular designs will increase in new technologies. Therefore, the benefits of regularity will increase with technology scaling.

REFERENCES

[1] "International Technology Roadmap for Semiconductors," 2012. [Online]. Available: http://public.itrs.net/Links/2012ITRS/Home2012.htm

[2] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no. 4, pp. 433–449, 2006.

[3] A. Asenov, "Random Dopant Induced Threshold Voltage Lowering and Fluctuations in Sub-0.1 (micron)meter MOSFET's: A 3-D 'Atomistic' Simulation Study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505–2513, 1998.

[4] L. T. Pang, "Measurement and Analysis of Variability in CMOS circuits," PhD, University of California at Berkeley, 2008.

[5] T. Ohtou, N. Sugii, and T. Hiramoto, "Impact of Parameter Variations and Random Dopant Fluctuations on Short-Channel Fully Depleted SOI MOSFETs With Extremely Thin BOX," *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 740–742, 2007.

[6] T. Matsukawa, S. O'uchi, K. Endo, Y. Ishikawa, H. Yamauchi, Y. X. Liu, J. Tsukada, K. Sakamoto, and M. Masahara, "Comprehensive analysis of variability sources of FinFET characteristics," in *VLSI Technology, 2009 Symposium on*, 2009, pp. 118–119.

[7] L. Pileggi, H. Schmit, A. J. Strojwas, P. Gopalakrishnan, V. Kheterpal, A. Koorapaty, C. Patel, V. Rovner, and K. Y. Tong, "Exploring regular fabrics to optimize the performance-cost trade-off," in *Proceedings of the 40th annual Design Automation Conference*, ser. DAC '03. New York, NY, USA: ACM, 2003, pp. 782–787.

[8] M. C. Smayling, H.-y. Liu, and L. Cai, "Low k1 logic design using gridded design rules," in *Design for Manufacturability through Design-Process Integration II*, V. K. Singh and M. L. Rieger, Eds., vol. 6925, no. 1. SPIE, Mar. 2008, pp. 69 250B–69 250B–7.

[9] J. Mauricio, F. Moll, and J. Altet, "Monitor strategies for variability reduction considering correlation between power and timing variability," in *24th IEEE International SOC Conference*. IEEE, 2011.

[10] M. Pons, F. Moll, A. Rubio, J. Abella, X. Vera, and A. Gonzalez, "VCTA: A Via-Configurable Transistor Array regular fabric," in *VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP DOI - 10.1109/VLSISOC.2010.5642683*, 2010, pp. 335–340.

[11] M. Pons, E. Barajas, D. Mateo, J. L. Gonzalez, F. Moll, A. Rubio, J. Abella, X. Vera, and A. Gonzalez, "Fast time-to-market with via-configurable transistor array regular fabric: A delay-locked loop design case study," in *2011 6th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*. IEEE, Apr. 2011, pp. 1–6.

APPENDIX C

PAPER 3: LOCAL VARIATIONS COMPENSATION WITH DLL-BASED BODY BIAS GENERATOR FOR UTBB FD-SOI TECHNOLOGY

Joan Mauricio and Francesc Moll.

Local Variations Compensation with DLL-based Body Bias Generator for UTBB FD-SOI technology

Joan Mauricio and Francesc Moll *Member, IEEE.*

***Abstract*— Local variations are increasingly important in new technologies. This paper presents the design of adaptive circuits based on the concept of Adaptive Body Bias Islands and a Forward and Reverse Body Bias Generator for FDSOI technology. The proposed Body Bias generator design, based on a DLL, allows a 70% area reduction compared to the DAC-based conventional design and reduces local variability by 85%. This closed loop implementation also allows to track runtime variations.**

***Index Terms*— Variability, Within-Die Variations, Sensors, Body Bias.**

I. Introduction

The semiconductor industry has based its progress on continuous transistor scaling, enabling faster chips with more functionalities, as well as reducing the unit cost of silicon devices. In recent years, portable devices have become prominent and power consumption is also becoming an increasingly important factor.

However, as scaled transistors enter the nanometer scale, manufacturing challenges increase substantially and give rise to undesired deviations in their parameters due to Process variations, thus altering their expected performance. In addition, environmental variations (Voltage and Temperature) increase the uncertainty in gate delay and power. Also Aging produces a parameter drift that affects differently the devices in the circuit. The main consequences of these PVTA variations are twofold: First, timing uncertainty requires an increase in safety margins that impair performance gains enabled by technology scaling. Second, power dispersion produces a decrease in parametric yield.

One way to counteract variability is to apply runtime adaptive mechanisms such that the circuit operates near nominal characteristics in the presence of static (Process) or even dynamic variations (Voltage, Temperature and Aging). One of the main adaptive mechanisms is the use of body bias to change threshold voltage [1]. When body bias is applied in a digital circuit it experiences substantial performance change, and hence nominal circuit performance can be recovered despite PVTA variations. Two types of body bias exist: Forward Body Bias (FBB) and Reverse Body Bias (RBB). FBB ($V_{BiasN} > gnd$, $V_{BiasP} < V_{Supply}$) reduces the transisor threshold and thus gate delay at the expense of an increase in leakage current.

The authors are with the High Performance IC Design Group, Dept. of Electronic Engineering, Universitat Politècnica de Catalunya.

This work was partly supported by the Spanish Ministry of Economy (MINECO) and ERDF funds through project TEC2013-45638-C3-2-R (Maragda).

RBB ($V_{BiasN} < gnd$, $V_{BiasP} > V_{Supply}$) increases threshold voltage and thus gate delay while leakage current diminishes.

The effectivity of body bias is known to decrease with nanometer bulk MOS devices. However, with the appearance of FDSOI technology with ultra thin body and buried oxide (UTBB), body bias is again seen as one of the best ways to control variability [2]. FDSOI allows a wide range of body bias voltages and consequently a larger control on the transistor threshold voltages.

Many papers have studied the application of body bias and the calculation of the required BB voltage, but few of them discuss practical implementation issues. The main contribution of the present paper is a novel design of a Forward and Reverse Body Bias Generator (FRBBG) circuit with a small area. The design is aimed to a variability adaptation scheme based on the concept of Body Bias Islands [3] that allows a fine-grain adaptivity.

The structure of the paper is as follows. Section II briefly reviews the concept of Adaptive Body Bias Islands to allow runtime adaptation to variability. In Section III it is presented the design of the FRBBG circuit. Section IV presents the simulation results of the FRBGG circuit. Finally, the conclusions are presented in Section V.

II. Adaptive Body Bias Islands

The variability problem appears not only from die to die, but also within die. Due to short-range process fluctuations, or local variations in voltage and temperature, differences exist between regions of an integrated circuit. Therefore, applying a single body bias voltage to the circuit does not avoid local differences. Instead, the die may be partitioned into a certain number of Body Bias Islands (BBI) with independent control of the Body Bias voltage.

The BBI approach was proposed before: [3], [4]. Also in [5] it is proposed a gate clustering approach that group gates with similar body bias requirement. In these proposals, some centralized algorithm is needed to assign the BB voltage for each BBI. Other proposals [6], [7] assume that not only Process variations have to be taken into account to reduce variability but also runtime variations such as voltage or temperature.

In the approach considered in this paper, each BBI has a sensor and a Body Bias Generator (BBG) with a self-adaptive procedure to adjust the body bias voltage at runtime so that the circuits contained in the BBI operate near to nominal conditions. The adaptation is local to each BBI and decentralized, so the overhead is only related to the sensor and BBG contained in the BBI.

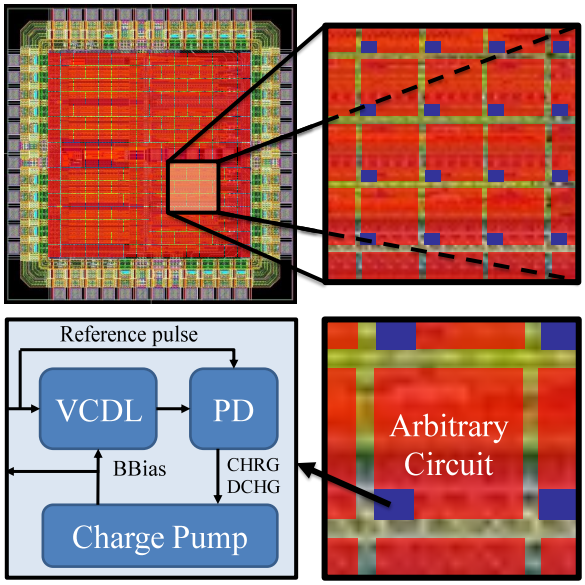

Fig. 1. Chip is partitioned into small body bias islands. Each island has a compensation circuit comprising a delay line, a phase detector and a charge pump.

Each body bias island (see Figure 1) contains a delay sensor, a phase detector and a body bias generator (BBG) based on a charge pump able to generate voltages both inside power supply limits (FBB) or outside this range (RBB).

III. Design of the BBG

Many BBGs have been proposed in several works [8], [9]. The main features proposed in this paper are the following: firstly, circuits are almost fully analog to reduce area penalty, power consumption and response time. This allows us to dispense with Digital to Analog Converters (DAC). Secondly, this BBG is able to generate both FBB and RBB which means that the circuit is able to generate voltages outside power supply rail. Moreover, we implement a sleep mode that puts the whole circuit at maximum RBB to reduce leakage. And finally, BBG works in closed loop based on a DLL-like structure so that the measured delay drift in each BBI is used to track variations by modifying body bias output. This fully distributed architecture fashion makes this proposal very scalable and easily integrable into an automated design flow.

The signal coming from an external pulse generator passes through a 32-stage Voltage Controlled Delay Line (VCDL) that, under zero body bias condition, will delay the reference pattern pulse according to PVTA variations. The reference signal pulse width corresponds to the nominal end-to-end delay of the VCDL in the absence of any variation for a given operating frequency. Hence, this BBG is compatible with dynamic frequency scaling techniques since the pulse width can be dynamically adapted to the operating clock frequency. The reference pulse frequency is set based on the desired BBG response time such that a faster response time implies a higher reference signal frequency at the expense of power consumption overhead.

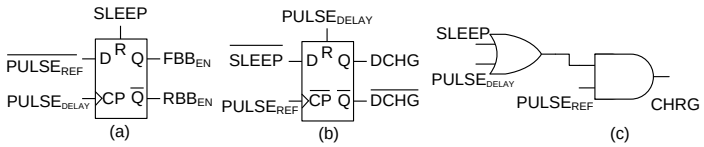

Fig. 2. Phase detector of the proposed FBB+RBB Generator.

Once the reference signal passes through the VCDL, both, delayed and reference signal arrive to the phase detector. This phase detector (see Figure 2) achieves its steady state when the VCDL output rising edge is produced at the same time as the reference signal falling edge, and thus VCDL nominal end-to-end delay (corresponding to the reference pulse width) is achieved. The phase detector generates both charge (RBB mode) and discharge (FBB mode) pulses as well as two signals that determine the BBG operation mode (RBB_{EN} or FBB_{EN}). The first flip-flop (Figure 2a) determines the operation mode: the reference pulse logic level is latched after each rising edge of the delayed pulse. If the VCDL is too fast, the inverse of the reference pulse will still be at low level, and therefore, RBB_{EN} will be active. RBB_{EN} will also be active in sleep mode in order to save leakage power. The flip-flop (Figure 2b) generates a discharge (*DCHG*) pulse (only when sleep mode is disabled) the width of which is proportional to the time difference between the falling edge of the reference pulse and the rising edge of the delayed reference pulse. Therefore, the slower the delay line, the wider the discharge pulse. Finally, the charge (*CHRG*) pulse is generated by making the AND2 operation between the reference pulse and its delayed version (Figure 2c). The charge pulse width will increase with VCDL speed, and hence charge pump will generate RBB. Also in sleep mode the phase detector will generate a charge pulse equivalent to the reference pulse, and thus RBB will reduce leakage in the target circuit.

Figure 3 shows the proposed charge pump for PMOS transistors body bias (the one for NMOS transistors is complementary to this) inspired in a two-stage Dickson charge pump [10]. This kind of charge pumps are commonly used as voltage multipliers and negative voltage supplies. In this proposal this circuit implementation is used to provide RBB voltages higher (for PMOS) and lower (for NMOS) than power supply voltage rails if needed. In addition to this feature, this circuit is also capable of generating FBB. When *CHRG* signal is at low level C1 is charged, and hence V_{C1} goes to $V_{DD} - V_T$. Afterwards, when *CHRG* signal goes to high level, V_{C1} doubles the high voltage level of the circuit and C2 is charged at $2(V_{DD} - V_T)$. Finally, when *CHRG* signal goes back to low level V_{C2} rises up to $3(V_{DD} - V_T)$. D1 and D2 prevent the discharge of C1 and C2 through V_{DD} and V_{C1} respectively. As *CHRG* and $\overline{CHRG}$ signal pulse widths depend on the phase error between the expected and measured delay on the VCDL, the wider the pulse, the faster the charge. When the VCDL achieves the expected delay, the circuit stops pumping charge and leaves the charge pump output (V_{BIASP}, V_{BIASN}) in steady state. The choice of capacitor values is a trade-off between area overhead, response time and BB voltage accuracy: the bigger

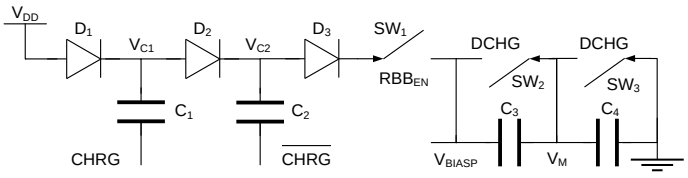

Fig. 3. The proposed FBB+RBB Generator implements a charge pump based in a Dickson Charge Pump.

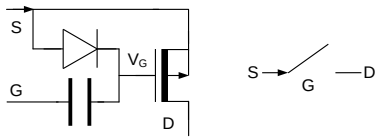

Fig. 4. Floating transmission gate schematic (left), and symbol (right).

the capacitors, the higher the accuracy and area overhead.

Only when VCDL is faster than expected (or during sleep periods) the RBB signal is high, and then the overvoltage generated by C1 and C2 charge a couple of capacitors (C3 and C4) connected in series, thus setting the BB voltage until steady state is reached. Otherwise, the floating transmission gate (see Figure 4) is open and the charge pump no longer charges the output capacitors. Due to the diode losses the maximum output of the circuit is around $2V_{DD}$, that is, an equivalent RBB of V_{DD}. These diodes are implemented by means of diode-connected NMOS transistors.

When the reference circuit is slower than expected, the phase detector disables RBB signal (SW_1 is opened) and thus C3 and C4 capacitors are no longer charged. What is more, two floating transmission gates slowly discharge these capacitors according to the phase shift between the delay line and the reference pulse width; the slower the delay line, the wider the *DCHG* pulse, and consequently, the faster the body bias decrease. As the maximum body bias voltage drop is $2V_{DD}$, at least two transmission gates in series are needed in order not to exceed Drain-Source Breakdown Voltage. Each of them discharges capacitors C3 and C4 respectively. As shown in Figure 4, these floating transmission gates are switched by means of a capacitive coupling since transistor gate is no longer referenced to ground (transistor source is at higher level). When gate terminal is set to high, the coupling capacitor produces a V_G voltage which $V_{DD} - V_T$ volts higher than the internal transistor source, and therefore, drives this transistor from cut-off to saturation. The main difference between floating transmission gates in Figure 3 is the W/L ratio of the internal transistor; while SW1 is expected to act as a switch (fast transistor, large ratio), SW2 i SW3 are expected to produce a small decrease in body bias voltage (slow transistor, small ratio).

IV. Simulation Results

A. Functional Simulation for the BBG

As observed in waveforms of Figure 5, at the beginning of the simulation the circuit is forced to enter the sleep state

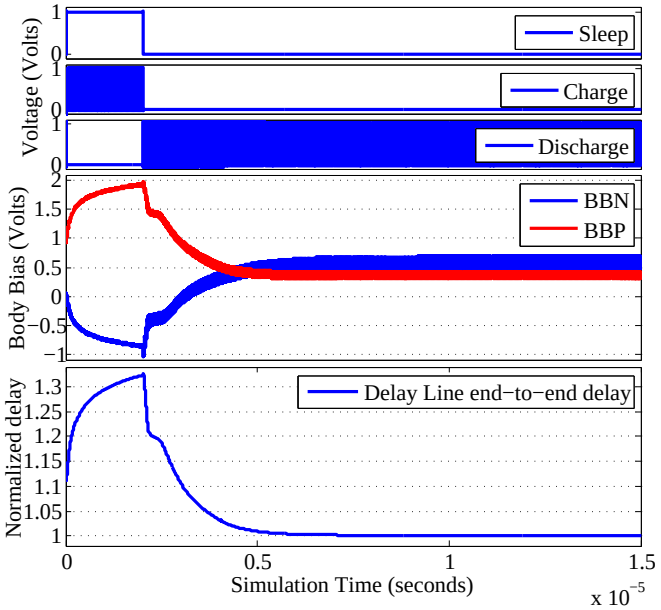

Fig. 5. Phase detector control signals (top), PMOS and NMOS body biases (middle) and VCDL end-to-end delay of the FBB+RBB Generator.

for 2 μs and RBB mode is enabled whatever it is the output of the phase detector. During sleep state the pattern pulse generator is used to charge the Dickson charge pump so that BBP and BBN get almost $2V_{DD}$ and $-V_{DD}$ respectively and, as a consequence, the BBI circuit including the VCDL transistors increase their V_{th}. This increase causes a reduction in leakage power consumption by a factor of 3.9 according to the simulations. It also produces a dramatic increase in the end-to-end delay, which is not relevant during sleep states.

Once the sleep state stops, the circuit tracks the input pattern pulse by discharging capacitors C3 and C4 (see Figure 3) and thus reducing BBP and BBN voltages until the end-to-end delay of the VCDL is equal to pattern generator pulse width (steady state). This occurs around 5 μs after sleep mode has been disabled.

B. Variability compensation

In order to test the effectiveness of this BBG circuit, 300 Monte Carlo simulations were performed using the foundry-provided variability data. In Figure 6 it can be seen the normalized end-to-end delay of the VCDL before (in blue) and after (in red) applying body bias. Before applying any adjustment delay samples are dispersed around the nominal delay with a coefficient of variation (understood as $3\sigma/\mu$) of 12.38%. After FBB+RBB is applied, the coefficient of variation is reduced to 1.83%. This residual variability can be explained by two main reasons. First, phase detector logic also suffers from variability, and therefore the same phase shifts between input signals will produce marginally different outputs. Second, dead zone effect in the phase detector produces an overall system phase inaccuracy.

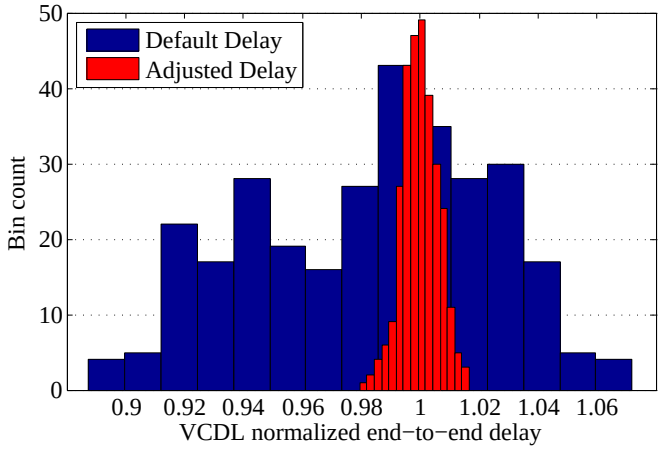

Fig. 6. VCDL end-to-end delay histograms (a.u.) with and without applying FBB+RBB voltage (300 Monte Carlo samples).

C. Comparison with other proposals

The area penalty has been estimated (layouts are not implemented yet) using the area of each component with a standard cell library for logic elements (in the phase detector and VCDL) and doing a first tentative placement for capacitors. Circuit components have been distinguished between transistors and capacitances. Transistor resources comprise the VCDL, the phase detector and the charge pump transistors. The main contribution to area penalty are the charge pump capacitances as they require up to 14 capacitors (320 μm^2) to generate both PMOS and NMOS body biases: 4 for RBB, 4 for FBB and 6 for the floating switches. As for the transistor resources it takes up 26 μm^2.

Power consumption overhead has also been computed. As this power consumption will depend on the external pattern generator frequency, simulations were done for different pattern frequencies and thus obtaining dynamic power consumption as a function of pattern frequency. As expected, dynamic power consumption overhead exhibits a linear dependency with the pattern frequency, being 38 μW per GHz.

A comparison chart of area overhead is presented in Table I. This table shows a comparison between the area of the proposed BBG in this article, and similar proposals in literature. In order to make a fair comparison between different technologies, circuit areas have been normalized (see Normalized area row) by dividing them by the square of the technology node and then, again, normalized to our circuit. Therefore, even assuming that circuits from other proposals would scale proportionally, the circuit proposed in this paper is significantly smaller. The main reason that explains this area improvement is the use of a VCDL in closed loop instead of DACs to generate BB.

V. Conclusions

In this paper a novel BBG design with FBB and RBB capability has been presented. This proposal exploits the body bias range extension that FDSOI allows in comparison with bulk CMOS, and therefore a larger control on the transistor threshold voltages is achieved. Simulations results show that this generator is able to reduce coefficient of delay variation from 12.38% to 1.83%, i.e. more than 6 times. Despite the use of closed loop is not a novel mechanism to control body bias [7], the main contribution of this paper is the charge pump design that not only replaces DACs as voltage generators (power and area savings), but is also capable to generate body bias voltages beyond supply voltage rails. Nevertheless, it has to be taken into account that body bias islands chip partitioning increases routing complexity since transistor bodies are no longer connected to power supply.

TABLE I
Comparison of the proposed BBGs with other proposals.

	This Proposal	**[9]**	**[8]**
Function	FBB+RBB	FBB+RBB	FBB
Process	28 nm	65 nm	90 nm
Circuit area	346 μm^2	5200 μm^2	30000 μm^2
Normalized area	1X	2.79X	8.39X
Power consumption	38 $\mu W/GHz$	600 μW	210 μW

References

[1] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 11, pp. 1396–1402, Nov 2002.

[2] P. Flatresse, B. Giraud, J. Noel, B. Pelloux-Prayer, F. Giner, D. Arora, F. Arnaud, N. Planes, J. Le Coz, O. Thomas, S. Engels, G. Cesana, R. Wilson, and P. Urard, "Ultra-wide body-bias range ldpc decoder in 28nm utbb fdsoi technology," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, Feb 2013, pp. 424–425.

[3] S. Garg and D. Marculescu, "System-level mitigation of WID leakage power variability using body-bias islands," in *Proceedings of the 6th IEEE/ACM/IFIP international conference on Hardware/Software codesign and system synthesis - CODES/ISSS '08*. New York, New York, USA: ACM Press, Jan. 2008, p. 273. [Online]. Available: http://www.researchgate.net/publication/221656795_System-level_mitigation_of_WID_leakage_power_variability_using_body-bias_islands

[4] ——, "System-level leakage variability mitigation for mpsoc platforms using body-bias islands," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 12, pp. 2289–2301, Dec 2012.

[5] S. H. Kulkarni, D. Sylvester, and D. Blaauw, "A statistical framework for post-silicon tuning through body bias clustering," in *Proceedings of the 2006 IEEE/ACM International Conference on Computer-aided Design*, ser. ICCAD '06. New York, NY, USA: ACM, 2006, pp. 39–46. [Online]. Available: http://doi.acm.org/10.1145/1233501.1233511

[6] N. Azizi and F. Najm, "Compensation for within-die variations in dynamic logic by using body-bias," in *IEEE-NEWCAS Conference, 2005. The 3rd International*, June 2005, pp. 167–170.

[7] R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "Mitigating parameter variation with dynamic fine-grain body biasing," in *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*, Dec 2007, pp. 27–42.

[8] M. Meijer, J. 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 *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, May 2010, pp. 2482–2485.

[9] N. Kamae, A. Islam, A. Tsuchiya, and H. Onodera, "A body bias generator with wide supply-range down to threshold voltage for within-die variability compensation," in *Solid-State Circuits Conference (A-SSCC), 2014 IEEE Asian*, Nov 2014, pp. 53–56.

[10] G. Palumbo and D. Pappalardo, "Charge pump circuits: An overview on design strategies and topologies," *Circuits and Systems Magazine, IEEE*, vol. 10, no. 1, pp. 31–45, First 2010.