# Robust and energy-efficient explicit pulse-triggered flip-flops in 28nm fdsoi technology for ultrawide voltage range and ultra-low power circuits

Sebastien Bernard

UNIVERSITÉ DE GRENOBLE

# THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 7 août 2006

Et de

## DOCTEUR DE L'UNIVERSITÉ CATHOLIQUE DE LOUVAIN

en **Sciences de l'Ingénieur**

Présentée par


## Sébastien BERNARD


Thèse dirigée par **Marc BELLEVILLE** et **Jean-Didier LEGAT,**
et co-encadrée par **Alexandre VALENTIAN** et **David BOL**

préparée au sein du **Laboratoire CEA-LETI**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal** de Grenoble
et dans **la Commission doctorale de domaine Science de l'Ingénieur et Art de bâtir et urbanisme** de Louvain

# Bascules à impulsion robustes en technologie 28nm FDSOI pour circuits numériques basse consommation à très large gamme de tension d'alimentation

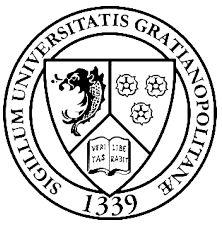
Thèse soutenue le **7 octobre 2014,**
devant le jury composé de :

**M. Marc BELLEVILLE**
Directeur de Recherche au CEA-LETI, co-Directeur de thèse
**M. Jean-Didier LEGAT**
Professeur à l'Université catholique de Louvain, co-Directeur de thèse
**M. David BOL**
Professeur-assistant à l'Université catholique de Louvain, Membre
**M. Alexandre VALENTIAN**
Ingénieur de recherche au CEA-LETI, Membre
**M. Andrei VLADIMIRESCU**
Professeur à l'ISEP, Paris, Rapporteur
**M. Wim Dehaene**
Professeur à la Katholeit Universiteit van Leuven, Rapporteur
**M. Lionel TORRES**
Professeur à l'Université de Montpellier II, Président

# ROBUST AND ENERGY-EFFICIENT EXPLICIT PULSE-TRIGGERED FLIP-FLOPS IN 28NM FDSOI TECHNOLOGY FOR ULTRA-WIDE VOLTAGE RANGE AND ULTRA-LOW POWER CIRCUITS

**Sébastien Bernard**
Université de Grenoble
Ecole Doctorale EEATS
and
Université catholique de Louvain
Ecole Polytechnique de Louvain


**Université catholique de Louvain**
**Louvain-la-Neuve (Belgium)**




**Université de Grenoble**
**Grenoble (France)**

*À mes parents et mon Papy*

# CONTENTS

# ACKNOWLEDGMENTS

Trois ans... En Belgique, tout le monde me disait "C'est beaucoup trop court, on commence seulement à être productif ! Mais au moins tu auras plus rapidement fini avec la thèse et tout ce qui va avec..." Et pourtant... Et pourtant j'ai vécu tant d'expériences pendant ces trois années, appris un nombre incalculable d'informations, vu, entendu et participé à tant d'évènements, que je n'ai pas l'impression que ces trois années soient passées si vite. J'achève maintenant une longue période de rédaction et de préparation de présentation et me jette dans la vie active, avec le sentiment que ces trois ans ne furent ni trop longs, ni trop courts, juste une expérience professionnelle, mais surtout humaine, riche et inoubliable.

Tout d'abord, je souhaiterais remercier mes encadrants, le Professeur David Bol et le Docteur Alexandre Valentian, qui ont toujours été disponibles lorsque j'avais des doutes ou des interrogations, ou lors de la relecture d'articles scientifiques. Même si la distance n'aidait pas toujours, ils ont toujours fait le maximum pour m'apporter une aide optimale. Plus particulièrement, je remercie David Bol pour m'avoir donné le goût et l'envie de la microélectronique à très haute efficacité énergétique, de par ses présentations passionnées et ses cours scientifiquement et esthétiquement remarquables, et Alexandre Valentian pour m'avoir accueilli et veillé à ce que je me sente chez moi au LISAN tout au long de ces trois années. Ensuite, j'aimerais remercier mes promoteurs, Marc Belleville et le Professeur Jean-Didier Legat, qui m'ont permis d'entamer cette thèse à Grenoble grâce à leur contact réciproque et leur appui de ma candidature au siège du CEA. Enfin, je remercie les autres membres de mon jury, les Professeurs Wim Dehaene, Andrei Vladimirescu et Lionel Torres pour les discutions riches et intéressantes qui ont suivi mes soutenances.

J'aimerais ensuite remercier mes collègues de bureau pour tout ce temps passé ensemble et pour avoir contribué à la bonne ambiance qui animait chacune de mes journées de travail : A Grenoble, Ogun Turkyilmaz, pour son apprentissage conventionnel du français non-conventionnel, Thiago Raupp Da Rosa, et son histoire après le match Allemagne-Brésil (mythique!), Santhosh Onkaraiah, pour son éternelle bonne humeur et son aide lors de l'organisation du voyage à Berlin, Olivier Billoint, pour avoir partagé tant d'abattements face aux nouveaux DK de ST, Sébastien Thuries, pour ses réponses patientes à mes questions sur les outils, Romain Lemaire pour son effarement quand il a appris ce qu'était le Mudday, Karim Azizi-Mourier, Youcef Fellah et Christian Servière du support informatiques, qui connaissaient tous mon prénom après seulement deux semaines, - et un remerciement tout particulier à Christian pour ces soirées

astronomiques en Chartreuse et cette magnifique Saturne!- l'équipe mémoire, et ses réunions presque biquotidiennes, composée entre autres de Adam Makosiej, et ses réponses patientes à mes nombreuses questions pas toujours intelligentes, et de Bastien Giraud, alias le marseillais-avec-ponctuation, et enfin mon équipe du Mudday, Alin Riatu, Matthieu Verdy et Robert Polster, pour cette expérience intense et inoubliable (la boue, la boue, la boue !). De l'autre côté à Louvain-la-Neuve, Guerric de Streel, surnommé l'amibe (n'en profite pas pour coloniser mon bureau !), avec qui je partageai conférence, cookies, recherche d'information dans un nuage de milliard de points, ainsi que joie et peine des simus, Vital Angelo Kuti Lusala, dont la patience face aux étudiants nous a tous impressionnés, Julien De Vos, qui découvrit Grenoble quand je retournais en Belgique, Geoffroy Gosset, pour ces match de tennis dignes des tournois du grand chelem, et enfin François Stas, pour son protocole de communication avec Guerric *per muros*.

Un immense merci à tous mes amis de l'association AITAP et mon coloc Benjamin Vianne, qui m'ont permis de rapidement m'intégrer et de participer à des activités sensationnelles ! Et pour terminer, je tiens tout specialement à remercier mes collègues grenoblois Bastien, pour l'argent de ses impôts qui a permis, à moi de partir en voyage aux quatre coins du monde et, à AITAP, de financer ses activités, et Robert, pour ce tandem franais-allemand qui restera dans les annales et qui nous a, j'en suis sr, tous les deux enrichis, pas seulement linguistiquement...

Je remercie le Centre à l'Énergie Atomique et aux Énergies Alternatives pour avoir financé mes recherches et mes séjours réguliers en Belgique.

Et finalement, je voudrais remercier tous les membres de ma famille qui m'ont soutenu pendant ces trois années de thèse, et tout spécialement mes parents qui ont parfois (souvent ?) stressé plus encore que moi, surtout lors de conférences tropicales durant lesquelles ils ont parfois eu du mal à dormir !

Sébastien

# ABSTRACT

The portable applications, such as smartphones, tablets, and mobiles, and the ultra-low power (ULP) circuits, such as RFID tags, wireless sensors network, and biomedical functions, are the applications driving the microelectronics industry today. In these applications, the microprocessor is connected to a battery or an energy harvesting system, meaning, in both cases, a finite amount of available energy and power. Therefore, the power and energy consumptions are of fundamental importance in the design of these circuits.

The FDSOI technology appeared recently in industrial processes to overcome the bulk technology limits and continue the trend of Moore's law. Thanks to the better electrostatic control of the channel, this technology provides a lower junction capacitance and leakage, steeper subthreshold slope, lower variability, and powerful back biasing technique over a wide voltage range. The back bias allows to dynamically modify the threshold voltage of the transistors in a reversible way. Therefore, this technology is extremely suitable for highly energy-efficient and ULP circuits.

In modern synchronous designs, the number of flip-flops (FF) has literally exploded with the raise of new microarchitectural techniques. Consequently, the flip-flop architecture has a decisive impact on the timing and energy consumption of the processor. The explicit pulse-triggered flip-flop (explicit pulsed-FF) topology presents remarkable timing properties which allow to gain a non-negligible part of the clock cycle. At the same time, its energy consumption can be severely reduced by sharing its pulse generator. So far, this structure is almost completely absent from circuits working at ultra-low voltage (ULV), where the master-slave architecture is mainly and widely used. The main reasons are its two following drawbacks:

- Low robustness against local variations in the pulse generation
- Positive hold time leading to additional delay buffers and thereby energy overhead

In order to improve the performances of UWVR and ULP circuits, both in a energy-efficiency and timing point of view, this dissertation studies and analyses architectural innovations to overcome these two disadvantages.

The study is driven by the two following questions:

- How to provide robust and energy-efficient pulse-triggered flip-flops at ultra-low voltage ?
- How can the FDSOI technology help us to improve the robustness and energy-efficiency ?

First, we present the explicit pulse-triggered flip-flop which is composed of a pulse generator (PG) sending pulse-like signals to a latch. We select the most promising latch architectures in an energy-efficiency point of view. Then, we compare these architectures in the energy-delay ($E - D$) domain by a sizing methodology at nominal and ultra-low voltages. The TGPL-Clk and $C^2$MOS-Data architectures are pointed out, respectively, for high-speed and low power applications. Afterwards, we show how the wide back biasing allowed in FD-SOI technology outperforms the sizing methodology in a delay and energy point of view. These comparisons are then confirmed by silicon measurement of the selected pulsed-FF architectures.

Second, we explain the robustness handicap of pulse-triggered flip-flop at ULV and the fundamental tradeoff between robustness and energy consumption. As a result, we propose a current-starved delay generator (DG) which provides sufficient degrees of freedom to reach the robustness target without impacting the energy consumption. Then, post-layout simulations and silicon measurements show the significant robustness improvement due to our proposed DG. The silicon measurements also show that choosing the proper back bias couple allows to reach the highest possible robustness. Afterwards, it is shown that additional flip-flop functionalities can be implemented in the PG with a very small timing, area, and energy penalties compared to master-slave architecture.

Finally, a conditional capture technique is presented to suppress the useless energy consumption remaining when there is no data activity. It is shown that, combined with the energy-efficient latch and PG previously analysed, this technique provides a lower energy consumption than master-slave architecture. After confirming the advantage of pulse generator sharing, the combination of previous innovations is integrated in a realistic register file. The comparison with master-slave based register files shows that our pulse-triggered flip-flops exhibit higher speed, lower area, and lower energy consumption over a wide range of supply voltage.

# ACRONYMS

| | |
|---|---|
| CDFF | Conditional Discharge Flip-Flop |
| CMOS | Complementary Metal-Oxide Semiconductor |
| $C^2$MOS | Complementary CMOS |
| $CP^3L$ | Conditional Push-Pull Pulsed Latch |
| CS | Current-Starved |
| DG | Delay Generator |
| FBB | Forward Body Bias |
| FDSOI | Fully Depleted Silicon-On-Insulator |
| FF | Flip-Flop |
| FinFET | Fin Field-Effect Transistor |
| FOM | Figure Of Merit |
| IC | Integrated Circuit |
| IoT | Internet-of-Things |
| MS | Master-Slave |
| PG | Pulse Generator |
| pulsed-FF | Pulse-triggered Flip-Flop |
| PVT | Process, Voltage and Temperature |
| P&R | Place&Route |
| RBB | Reverse Body Bias |
| SoC | System-on-Chip |
| SOI | Silicon-On-Insulator |
| TGPL | Transmission-Gate Pulsed Latch |
| ULP | Ultra-Low-Power |
| ULV | Ultra-Low-Voltage |

# LIST OF NOTATIONS

| | | |
|---|---|---|
| $\alpha_{sw}$ | Data switching activity factor | $[-]$ |
| $\alpha_{rate}$ | Input rate activity factor | $[-]$ |
| $\beta_{PN}$ | Ratio between the PMOS gate width and the NMOS gate width in the same CMOS branch | $[-]$ |
| $\sigma_{V_{th}}$ | Threshold voltage standard deviation | $[mV]$ |
| $C_g$ | Intrinsic gate capacitance | $[F]$ |
| $C_L$ | Load capacitance | $[fJ]$ |
| $C_{out}$ | Output filtering capacitance | $[F]$ |
| $C_{ox}$ | Gate oxide capacitance | $[F]$ |
| $D_0$ | FO4 inverter chain propagation delay at given PVT | $[s]$ |
| $E_0$ | Minimum sized symetrical inverter energy dissipation at given PVT | $[J]$ |
| $E_{dyn}$ | Dynamic energy per operation | $[J]$ |
| $E_{min}$ | Minimum energy per operation | $[J]$ |
| $E_{op}$ | Energy per operation | $[J]$ |
| $E_{stat}$ | Static energy per operation | $[J]$ |
| $f_{clk}$ | Clock frequency | $[Hz]$ |
| $gnd$ | ground (voltage reference) | $[-]$ |
| $gnds$ | NMOS back plane value or NMOS back bias | $[V]$ |
| $I_0$ | Subthreshold reference current | $[A/\mu m]$ |
| $I_{leak}$ | Circuit leakage current | $[A]$ |
| $I_{off}$ | Off-state drain current | $[A]$ |
| $I_{on}$ | On-state drain current | $[A]$ |
| $I_{sub}$ | Subthreshold drain current | $[A]$ |
| $L_g$ | Gate length | $[A]$ |

| | | |
|---|---|---|
| $n$ | Body-effect factor | $[-]$ |
| $P_{dyn}$ | Dynamic power consumption | $[W]$ |
| $P_{stat}$ | Static power consumption | $[W]$ |
| $P_{sc}$ | Short-circuit power dissipation | $[W]$ |
| $P_{sw}$ | Switching power dissipation | $[W]$ |
| $S$ | Subthreshold swing | $[mV/dec]$ |
| $U_{th}$ | Thermal voltage | $[mV]$ |
| $T_{clk}$ | Clock period | $[s]$ |
| $V_{BB}$ | Body bias voltage | $[V]$ |
| $V_{bs}$ | Body-to-source voltage | $[V]$ |
| $V_{dd}$ | Supply voltage | $[V]$ |
| $V_{dd,min}$ | Minimum operating supply voltage | $[V]$ |
| $V_{ds}$ | Drain-to-source voltage | $[V]$ |
| $V_{gs}$ | Gate-to-source voltage | $[V]$ |
| $V_{th}$ | Threshold voltage of MOS device | $[V]$ |
| $vdds$ | PMOS back plane value or PMOS back bias | $[V]$ |
| $W_g$ | Gate width | $[\mu m]$ |

# INTRODUCTION

Over the last two decades, the portable applications have become the keystone of the microelectronics industry. Millions of smartphones, tablets, and mobiles are sold every day and the projections do not see a decrease for a while. However, the energy bugdet is the bottleneck of these applications. Today, the smartphones need to be recharged every day and the battery has become a major concern for industrial people and customers. The most efficient solution found by designers is to reduce the supply voltage of the CMOS circuits. This greatly decreases each component of the energy dissipation, with the inconvenience of decreasing the maximum speed of the circuit. The current targeted tradeoff is to work at nominal voltage when high-performances are needed, for example when a webpage is downloaded, and at ultra-low voltage when performance is not on concern. Consequently, the supply voltage of these circuits covers a wide range of value during the life-time of the circuit. It is why there are called ultra-wide voltage range (UWVR) circuits [1].

At the same time, emerging applications such as biomedical devices, wireless sensors networks, radio-frequency identification (RFID) tags, and the advent of the Internet of Things (IoT) paradigm have led designers to develop ultra-low-power (ULP) design of integrated circuits [2]. Most of these circuits will work with systems harvesting only the available energies in the environment. Thus, the energy budget will determine the accomplishment of these circuits. In the IoT, wireless sensors will be placed outdoor and indoor and biomedical measurement tool will be on- and then in-body. Thereby, only a small amount of energy and power will be available in the environment. Therefore, there is a huge demand for reducing the energy consumption of the circuits dedicated to portable and ULP applications.

To continue the trend of Moore's law [3], new technologies have emerged to replace the conventional bulk technology. Two of them are today implemented in industrial process flow: the FinFET technology and the fully-depleted silicon on insulator (FDSOI) technology. The aim of these two transistor technologies is a better electrostatic control of the channel. The FinFET transistor has a 3D shape which allows to encircle the channel, and the FDSOI transistor presents a buried oxide (BOX) below the channel which acts like a second gate and brings many others advantages. In this work, the transistor architectures are designed

and fabricated in 28nm FDSOI technology, with full benefit of its interesting properties.

The appropriate choice of flip-flop architecture is of fundamental importance in the design of VLSI integrated circuits. In modern synchronous CMOS circuits, the clock tree and its leaves represent between 30% and 70% of the total energy consumption of the microprocessor [4, 5]. The main reason is that the number of flip-flops has literally exploded the last two decades. Pipelining, super-scalar and time-borrowing techniques need always more flip-flops to reach the timing limits of the circuit. Therefore, the flip-flop timing and energy characteristics have a considerable impact on the performances of the whole circuit.
As a result, this dissertation starts with an overview of the flip-flop topologies with the aim of selecting the most promising FF structure for improving the energy-efficiency of UWVR and ULP circuits. It turns out that explicit pulse-triggered flip-flops own remarkable properties for the targeted applications, but present several disadvantages to function at ultra-low voltage. The pulse-triggered flip-flop (pulsed-FF) is a well-known topology which is, to the best of our knowledge, only used in high-performances circuits. It is composed of one latch, which is open during a short period determined by a pulse-like signal. This signal is generated by a pulse generator active at the triggering clock edge.

Unfortunately, this flip-flop topology suffers from two big drawbacks compared to master-slave architecture, which are magnified at ultra-low voltage: it presents a lower robustness to local variations ; it exhibits a positive hold time, inducing energy overhead paid for additional delay buffer insertion at Place&Route step.

This is the focus of this dissertation: integrating the fast pulse-triggered flip-flop topology in energy-efficient circuits working at ultra-low voltage operations, with the help of the FDSOI technology.

## I.1 THE FDSOI TECHNOLOGY

The fundamental configuration of the FDSOI transistor is a conventional bulk transistor where a thin oxide layer, or buried oxide (BOX), is inserted between the substrate and the active part such as the channel height is a few nanometre (8nm of silicon for 28nm technology for a BOX of 25nm height). Beneath the BOX, the region is called the back plane and is not necessarily tied to the supplies. This topology presents many advantages. First, the junction capacitance between the source-drain contact and the bulk is reduced, as well as the junction leakage current. Then, the better electrostatic control provides a dramatic reduction of many short channel effects and an enhanced subthreshold slope. Furthermore, FDSOI technology exhibits a much lower variability than bulk technology mainly thanks to the undoped transistor channel. Moreover, it is possible to produce up to three different threshold voltages without doping the channel and adding variability. And as a last advantage, the most important property of the FDSOI technology according to the author, the back biasing technique. The voltage below the buried oxide, called back bias, can vary over a

wide range of voltage. The variation is equivalent to a strong forward (FBB) or reverse body bias (RBB) in bulk.

## I.2 PURPOSE OF THIS THESIS

This dissertation studies and develops pulse-triggered flip-flops with the aim of pushing them in ultra-low voltage operations. In addition to the robustness challenge, ULV operations are focusing on the energy-efficiency of the circuits. In order to explain the targets followed all over this work, let us remind that the energy-delay product (EDP) is the main figure of merit for UWVR circuits, while the energy per operation ($E_{op}$) is of primary importance in ULP applications.

Architectural innovations are presented and designed in this thesis to answer the following questions:

- What is the most energy-efficient latch architecture to be inserted in an industrial standard-cell library covering a wide range of supply voltage ?

A lot of studies [6, 7, 8, 9] compare flip-flop architectures at the same, and often nominal, supply voltage, trying to reach a given figure of merit (delay, energy, or energy-delay product). Moreover, the compared flip-flops hardly present more than three functional pins (input, output, and clock) which is not realistic for applications in advanced technologies. In this work, we elaborated a set a promising scannable and resettable flip-flop topologies and then compared them in the energy-delay ($E - D$) domain at nominal and ultra-low supply voltage. Afterwards, we studied the back biasing technique applied to the selected architectures and highlighted the benefit in timing and energy performances provided by a wide back bias range.

- How to guarantee the robustness of pulse-triggered flip-flop facing local variations at ultra-low voltage, without overdesign and energy overhead ?

As the width of the pulse window is determined by a delay generator composed of a chain of CMOS stages, the generated delay is highly impacted by local variations at ultra-low voltage. If the pulse width is too narrow, the latch of the pulsed-FF does not have enough time to change its state. The common ways for ensuring the minimum pulse width is wide enough under worst-case conditions, lead to an overdesign and/or an significant energy overhead. In this work, we proposed a current-starved delay generator as architectural innovation to overcome the tradeoff between robustness and energy and studied how to choose the back bias values to maximise the yield under timing constraints.

- What is the optimal usage of pulse-triggered flip-flops at block level, taking into account the synthesis, placement and routing constraints ?

After the robustness, the second main drawback of pulsed-FF is its positive hold time which leads to additional delay buffer insertion, thus energy consumption, during synthesis and placement. We proposed a new pulse generator with

a pseudo-XOR gate inside to make the hold time negative and limit the energy consumption when no data activity. Then, we assembled several innovations presented in the dissertation to lay out a robust and energy-efficient register file and to show the obtained gains in speed, energy, and area.

## I.3 THESIS OUTLINE

In order to answer the previously exposed questions, this thesis is organised as follows:

*Chapter 1.* After a short reminder about the power and energy consumption of digital CMOS circuits, the FDSOI technology and all its advantages are presented in more details. Next, we assess the state of the art of the four CMOS flip-flop topologies. The master-slave, differential, pulse-triggered and dual-edge configurations are illustrated with FF architectures from the literature and, for each of them, the advantages and disadvantages are exposed and finally summarized. At the end, we point out the explicit pulse-triggered flip-flop topology as a promising candidate to increase the energy-efficiency of the UWVR and ULP circuits. This topology suffers from two main drawbacks, namely the robustness and the positive hold time, which are handled in Chapter 3 and 4 respectively.

*Chapter 2.* After having observed all the architectural ideas in the previous chapter, we perform a design reasoning to select the most efficient flip-flop architectures for the targeted applications. This leads us to compare six pulse-triggered architectures. The comparison consists of determining the set of design points, *i.e.* transistor gate width, which provides optimal points in the energy-delay ($E-D$) domain, from the high-speed region to the low power region. The pulsed-FFs architectures are compared in the $E-D$ domain both at nominal and ultra-low supply voltages. While the transmission gate pulse latch muxed clock (TGPL-Clk) architecture presents the best energy-efficiency for high-speed operations, the complementary CMOS muxed data ($C^2$MOS-Data) architecture is revealed as the most energy-efficient pulsed-FF architecture over a wide range of targeted delays and supply voltages. Afterwards, we show how the back biasing technique allowed by the FDSOI technology, can provide better energy and delay performances than the sizing methodology used before. Finally, silicon measurements confirm the results obtained previously.

*Chapter 3.* This chapter starts by explaining the inherent tradeoff of pulse-triggered structures at ULV. To ensure a sufficient robustness, one of the two main drawbacks of pulsed-FFs, a large energy overhead is paid in several ways. To overcome this issue, we propose a current-starved delay generator (DG) presenting enough degrees of freedom to reach the desirable robustness without energy penalty. Then, post-layout simulations and silicon measurements show how our proposed DG structure improves dramatically the robustness of pulsed-FFs.

Finally, we present a way of implementing flip-flop additional functionalities, usually added in standard-cells library, in the pulse generator, thus providing a robust and energy-efficient pulse-triggered flip-flop.

*Chapter 4.* Firstly, the behaviour of pulsed-FFs face to the clock skew appearing at clock tree synthesis, is exposed. Afterwards, a new conditional capture technique is presented and explained. As the pulse generator (PG) is the highest energy consumer in pulsed-FF architectures, this structure efficiently tackles the FF energy dissipation. Furthermore, the $E_{op}$ is reduced up to the point that the pulsed-FF finally exhibits a lower energy consumption than master-slave architecture. Then, after confirming the efficiency of the shared pulse generator, we integrate previous innovations, namely fast and energy-efficient latch, robust and energy-efficient DG and shared PG, in a structured register file. This pulsed-FF based register file presents a higher speed, lower energy consumption and lower area compared to master-slave based register files.

*Conclusions and appendixes.* We finally summarize this work and draw some perspectives. Additionally, the conclusion is followed by 3 appendixes. In Appendix A, additional studies about pulsed-FF architectures are exposed. The optimal conditional discharge flip-flop (CDFF) structure version is justified, regarding the precharge mechanism and the reordering technique, and the inefficiency of the resettable and scannable conditional precharge flip-flop (CPFF) is exposed. In Appendix B, the testbench used for the flip-flop comparison in Chapter 2 is presented in details. In particular, the delay and energy are rigorously defined. Finally, Appendix C exposes the testbench used for the register file comparison of Chapter 4.

*Note.* Let us highlight that the basic edge-triggered sequential element is sometimes called D-latch in the literature, mostly american. It will be called *flip-flop* in the rest of this work and *latch* will be used for level-sensitive elements.

# AUTHOR'S PUBLICATION LIST

---

### Related journal papers

JP1. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "A Robust and Energy Efficient Pulse-Triggered Flip-Flop Design for Ultra Low Voltage Operations", Journal of Low Power Electronics, Vol.10, pp. 1-9, 2014.

JP2. E. Beigne, R. Wilson, P. Flatresse, A. Valentian, F. Abouzeid, T. Benoist , C. Bernard, S. Bernard, O. Billoint, S. Clerc, B. Giraud, A. Grover, J. Le Coz, I.M. Panades, J.-P. Noel, B. Pelloux-Prayer, P. Roche, O. Thomas, Y. Thonnart, D. Turgis, F. Clermidy, P. Magarshack, "A 460MHz at 397mV, 2.6GHz at 1.3V, 32b VLIW DSP Embedding FMAX Tracking Techniques for PVT Variations Compensation", IEEE Journal of Solid-State Circuit (JSSC), solicited and accepted.

### Related conference papers

CP1. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "An efficient metric of setup time for pulsed flip-flops based on output transition time", International Conference on Integrated Circuit Design and Technology Proceedings (ICICDT), pp. 9-12, 2013.

CP2. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "A Robust and Energy Efficient Pulse Generator for Ultra-Wide Voltage Range Operations", Proceedings of the 5th Asia Symposium on Quality Electronic Design (ASQED), pp. 80-84, 2013.

CP3. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "Design of a Robust and Ultra-Low-Voltage Pulse-Triggered Flip-Flop in 28nm UTBB_FDSOI Technology", SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), pp. 1-2, 2013.

CP.4 E. Beigne, A. Valentian, B. Giraud, O. Thomas, T. Benoist, Y. Thonnard, S. Bernard, G. Moritz, O. Billoint, Y. Maneglia, P. Flatresse, J.P. Noel, F. Abouzeid, B. Pelloux-Prayer, A. Grover, S. Clerc, P. Roche, J. Le Coz, S. Engels, R. Wilson,, "Ultra-Wide Voltage Range Designs in Fully-Depleted Silicon-On-Insulator FETs", Design, Automation & Test in Europe (DATE), 2013.

CP.5 R. Wilson, E. Beigne, P. Flatresse, A. Valentian, F. Abouzeid, T. Benoist , C. Bernard, S. Bernard, O. Billoint, S. Clerc, B. Giraud, A. Grover, J. Le

Coz, I.M. Panades, J.-P. Noel, B. Pelloux-Prayer, P. Roche, O. Thomas, Y. Thonnart, D. Turgis, F. Clermidy, P. Magarshack, "A 460MHz at 397mV, 2.6GHz at 1.3V, 32b VLIW DSP, embedding FMAX tracking", the Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, pp. 452-453, 2014.

CP.6 S. Bernard, A. Valentian, M. Belleville, J.D.-Legat, and D. Bol, "Experimental Analysis of Flip-Flops Minimum Operating Voltage in 28nm FD-SOI and the Impact of Back Bias and Temperature", Power and Timing Modeling, Optimization, and Simulation (PATMOS), 2014.

### Unrelated papers

UP1. D. Bol, S. Bernard, D. Flandre, "Pre-Silicon 22nm Compact MOSFET Models for Bulk vs. FD SOI Low-Power Circuit Benchmarks", Proceedings of the SOI Conference, 2011.

UP2. F. Botman, J. De Vos, S. Bernard, J.D. Legat, D. Bol, "Bellevue: a 50MHz Variable-Width SIMD 32bit Microcontroller at 0.37V for Processing-Intensive Wireless Sensor Nodes", in IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1207-1210, 2014.

### Invited tutorials and keynotes

ITK1. D. Bol, J. De Vos, F. Botman, G. de Streel, S. Bernard, D. Flandre, and J.-D. Legat, "Green SoCs for a Sustainable Internet-of-Things", in *Proc. Workshop Faible Tension Faible Consommation (FTFC)*, 4 p., 2013.

ITK2 D. Flandre, O. Bulteel, G. Gosset, P.-A. Haddad, S. Bernard, B. Rue, D. Bol, "Disruptive ultra-low-leakage design techniques for ultra-low-power CMOS circuits", in *Proceedings of the CMOS Emerging technologies Conference*, 2012.

### Workshop

W1. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "An Energy Efficient Pulse-Triggered Flip-Flop Robust to Local Variations in Ultra Low Voltage", VARI 2013.

**CHAPTER 1**

# CONTEXT, BACKGROUND AND MOTIVATION

**Abstract**

This chapter starts with a discussion about the consequences of the technology scaling on the performances of digital CMOS circuits, showing the limits of the conventional bulk technology. Considered as a major solution to overcome these problems, the fully-depleted silicon on insulator (FDSOI) technology is presented. The benefits of this disruptive technology are described and new design possibilities are highlighted. All that motivates the choice of the FDSOI transistor technology for the high-speed and low-power applications.

Afterwards, a brief look at the current microprocessor architectures exhibits the importance of the flip-flop element. Then, a large overview of the flip-flop state of the art architectures is presented. Different topologies, their characteristics, and their performances in the figures of merit of flip-flops are studied and compared to fully understand the advantages and disadvantages of each one. As a result of this qualitative comparison, we motivate the line of investigation chosen in this thesis : the explicit pulse-triggered architecture is suggested as the best candidate for developing a fast and energy-efficient flip-flop. The current limitations and drawbacks of this topology are mentioned and will then be tackled in the next chapters.

**Contents**

## 1.1 INTRODUCTION

For decades, the main target of the microelectronics industry has been the increase of the speed of digital circuits while reducing the fabrication cost. This had been reached by two main ways of improvement: the scaling of the transistor dimensions and the architectural optimisation of the processor. Technology scaling has allowed to increase the speed and the integration density of digital circuits while reducing the overall power consumption. Nevertheless, we are today reaching the limits of this trend both from a technological and a design point of view. As the dimensions scale down, the variability of electrical properties becomes more and more predominant, leading finally to an unacceptable margin and energy overhead [2]. Many disruptive technologies have been proposed to replace the conventional bulk, especially the fully-depleted silicon on insulator (FDSOI) technology. This technology provides speed and energy gains with minimum change in the technological process. The section 1.2 explains how the addition of a buried oxide layer provides a great improvement in the transistor properties and how it can help designer to improve the energy-efficiency.

In the same time, the architecture of the central processing unit (CPU) has drifted from a single instruction by clock cycle to a deeply pipelined and also multi-scalar CPU architecture. Pipeline technique adds several registers along the datapath in order to subdivide instructions into many stages [10]. Multi-scalar architecture is a partial replication of the datapath allowing the simultaneous processing of N data for an approximate cost of N in hardware [10]. Therefore, while heavily increasing the throughput, these architectures need a largely superior amount of sequential elements leading the entire clock tree to become the biggest power consumption part of a modern digital circuit [4, 5, 6]. As a result, it is of primary importance to choose the proper flip-flop architecture in order to provide fast and energy-efficient processors.

The chapter is structured as follow: section 1.2 gives an overview of the trend of the microelectronics industry and why it has reached its limits with the classical bulk technology. Then, the Fully-Depleted Silicon On Insulator (FDSOI) technology is presented as well as its capabilities to help designers to overcome the limits of scaling. Afterwards, section 1.3 briefly introduces the modern processor architectures and exhibits the importance of flip-flop (FF) architecture on the speed and energy of sequential digital circuits. An overview of the state-of-the-art flip-flop architectures is performed for the four big categories of FF: master-slave, differential, pulse-triggered and dual edge. The section concludes by pointing out the remarkable properties of explicit pulse-triggered flip-flop (pulsed-FF) topology and motivates the research approach of the following chapters.

## 1.2 TECHNOLOGICAL CONTEXT

From a technological point of view, the reduction of the fabrication cost was achieved by the trend of Moore's law: the transistor density in a microprocessor doubles every 18 to 24 months. The reduction of transistor dimensions has for direct impact to decrease, for the same amount of functionalities, the area of the circuit and thus the cost per chip per wafer. This last point is obviously a determinant parameter for an industrial point of view and that is why this trend is continuing. This section explains why this law is also interesting for two other important figures of merit of a microelectronic circuits - the speed and the energy consumption - and why this trend is reaching its limits with the advanced technology. Finally, the FDSOI technology and its new design facilities will be presented.

### 1.2.1 Delay and energy equations

Before explaining the impact on the circuit performances, let us remind the fundamental equations giving the speed and the energy consumption of a digital circuit. It allows to understand the evolution of these performances with the technological parameters.

The propagation delay of a CMOS logic gate is proportional to:

$$T_{del} \propto C_L \frac{V_{dd} - V_{th}}{I_{on}} \tag{1.1}$$

where $C_L$ is the typical load capacitance of the transistors, $V_{dd}$ the supply voltage, $V_{th}$ the MOSFET threshold voltage, and $I_{on}$ the average MOSFET drain current in on-state.

The instantaneous power of a digital circuit is composed of two components: the dynamic power, only consumed when circuit performs computation, and static power.

$$P_{inst} = P_{dyn} + P_{stat} \tag{1.2}$$

The static power can be expressed as

$$P_{stat} = I_{leak} V_{dd} \tag{1.3}$$

where the leakage current $I_{leak}$ depends exponentially on the threshold voltage $V_{th}$ [2]:

$$I_{leak} \propto e^{\frac{V_{gs}+\eta V_{ds}}{S}} \times (1 - e^{\frac{-V_{ds}}{U_{th}}}) \tag{1.4}$$

where $\eta$ is the DIBL coefficient, S is the subthreshold swing equal to $ln(10)nU_{th}$, $n$ the body-effect factor and $U_{th}$ the thermal voltage.

The dynamic power also comes from two components: the capacitance switching power and the short circuit current. The capacitance switching is the power

needed to charge and discharge the capacitance of the circuit, especially the transistor gate capacitances and interconnection wires.

$$P_{sw} \propto C_L V_{dd}^2 f_{clk} \tag{1.5}$$

where $f_{clk}$ is the clock frequency. If the circuit is properly sized, the input slope on the gate is sufficiently abrupt so that the short circuit power $P_{sc}$ can be kept at 5-10% of $P_{sw}$ [11].

The energy per operation can be extracted by integrating the power over the clock cycle:

$$E_{tot} = E_{dyn} + E_{stat} \tag{1.6}$$
$$E_{tot} \approx E_{sw} + E_{stat} \tag{1.7}$$
$$E_{tot} \propto C_L V_{dd}^2 + I_{leak} V_{dd} T_{clk} \tag{1.8}$$

where the short circuit energy is neglected in equation 1.7 and $T_{clk} = 1/f_{clk}$ is the clock period.

Reducing transistor size has led to a reduction of the supply voltage in order to guarantee the same electrical field across the gate dielectric and to avoid an electrical breakdown. The threshold voltage is reduced to keep the same overdrive voltage and thus the current in on-state. The gate delay was thus reduced because both $C_L$ and $V_{dd}$ are reduced while $I_{on}$ remained roughly the same. Although a decrease of the threshold voltage increases the leakage current, the overall energy was effectively reduced because the leakage energy stayed quite low compared to dynamic energy. Therefore, reducing dimension and thus supply voltage permitted the microelectronics industry to create digital circuits with lower cost, higher speed, and lower energy consumption.

Now, let us have a look on the last statement for ULV and ULP circuits. As the clock period depends on the delays of the gates ($T_{del}$) and both $I_{on}$ and $I_{leak}$ depend on supply voltage, equation 1.8 has a non-trivial dependence on $V_{dd}$. Actually, the shape of this function exhibits a global minimum at a given $V_{dd}$ [2]. From an energy point of view, two $V_{dd}$ are concerned: the $V_{dd,opt}$ is the supply voltage which gives the lowest energy per operation. If a circuit has a certain amount of computation to do before coming back in stand-by mode, it should operate at this supply voltage to minimize the amount of energy consumed by computation. The $V_{dd,min}$ is the minimum supply voltage which guarantees the correct functionality of the circuit. If a circuit has to permanently compute data and is concerned about power, it should work at this voltage to minimize its power consumption.

Those two voltages are highly impacted by the transistor variability, as explained in the next section.

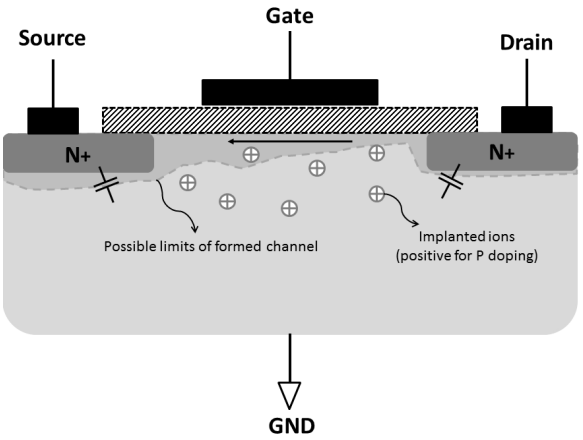



**Figure 1.1:** Conventional Bulk transistor (NMOS).

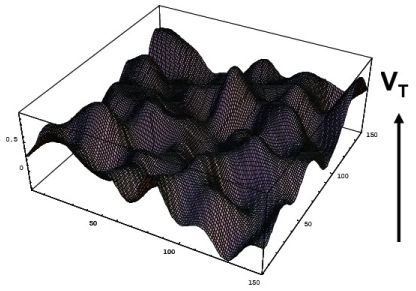



**Figure 1.2:** Discrete threshold voltage distribution in the transistor channel [13].

### 1.2.2 The end of scaling benefit in bulk technology

The conventional bulk transistor is represented in figure 1.1. The electrical relations between currents and voltages depend on many technological parameters, such as the effective gate width, gate length, the oxide thickness,... and the doping concentration. All of those parameters are slightly modified during the fabrication process, leading to a variability in voltage-current relation between the transistors on a same chip and on the whole wafer.

Scaling down and down the technology, the size of the transistors becomes so small that the discrete number of the implanted ions and the roughness of the layer interfaces have a huge impact on the average threshold voltage. It has been shown that the critical device parameter variability is the threshold voltage variability, mainly caused by the doping statistics in the channel region for bulk technology [12]. Figure 1.2 illustrates the effect of the dopant position on the threshold voltage.

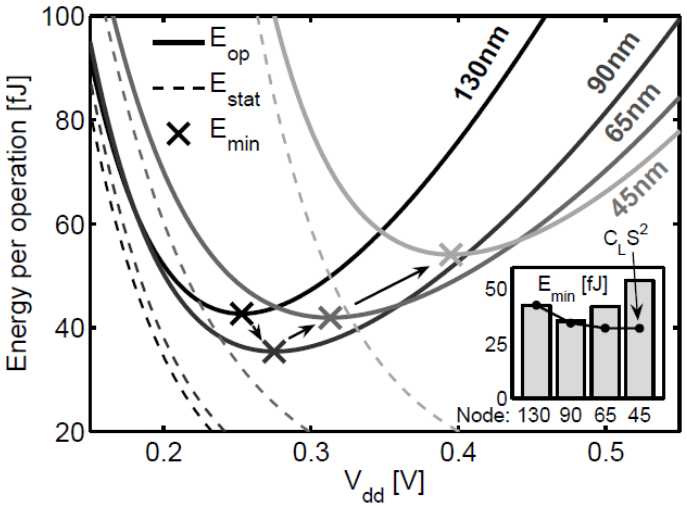



**Figure 1.3:** Minimum energy point increases with further technological node because of the higher variability [2].

To study the evolution of performances with the scaling, the effects of the variability must be taken into account. To do that, we have to introduce an equation which has not been mentioned in the previous section, the Pelgrom's law. It states that the variability of the threshold voltage of transistors among the whole chip follows a Gaussian distribution with standard deviation

$$\sigma_{V_{th}} = \frac{A_{V_{th}}}{\sqrt{W_g L_g}} \tag{1.9}$$

where $W_g$ and $L_g$ are respectively the transistor gate width and length and $A_{V_{th}}$ a parameter depending on the technology.

Last decade, the variability induced by the aggressive sizing of the transistor led designers to take huge margins to ensure the correct functionality of the chip. As shown in figure 1.3 and thoroughly explained in [2], the minimum energy point is no longer lowered by reducing the size of the technological node. Thus, because of the variability, the energy consumption of the advanced technological nodes is actually bigger, in contrast to the basics laws of scaling. On top of that, the variability has a direct impact on $V_{dd,min}$ [14].

To overcome this limit in Moore's law, new transistor technologies were suggested in the last decade. Today, the most competitive technologies are the finger field-effect transistor (FinFET) and the fully-depleted silicon on insulator (FD-SOI).

In FinFET technology, also called trigate in certain companies, the transistor channel is surrounded on three sides by the same poly-silicon layer. It was developed and presented to push still away the speed of the high-performance

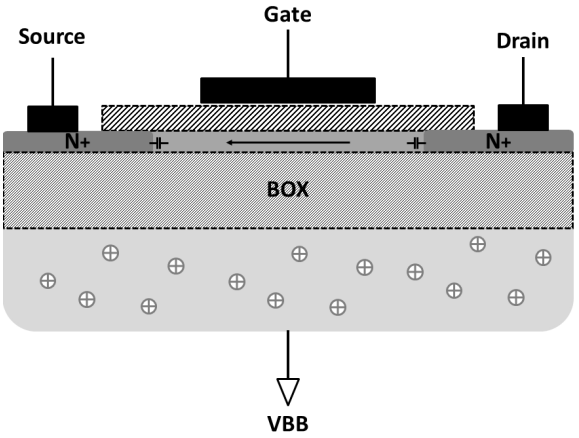



**Figure 1.4:** FDSOI transistor (NMOS version).

circuits [15, 16]. On the other hand, the FDSOI technology [17, 18, 19] has been presented as a very promising candidate for low power applications and will be described in the next section.

### 1.2.3 FDSOI transistor technology

The FDSOI transistor is represented in figure 1.4. A thin oxide layer, or buried oxide (BOX), is inserted between the bulk and the active part such as the channel height is a few nanometer (8nm of silicon for 28nm technology for a BOX of 25nm height). Beneath the BOX, the bulk region is now called the back plane. The back plane might be different to the bulk of the wafer and is not necessarily clamped to the supplies. When the gate-source voltage becomes higher than the threshold voltage, the electrons coming from the drain and source completely fill the channel between the gate and thick oxide insulator. The silicon of the electron channel is thus entirely in deep depletion regime, giving the name of the technology : Fully-Depleted Silicon on Insulator.

As first observation, the junction capacitance between the source and drain and the bulk is now reduced to a little PN junction within the channel. It provides a junction capacitance per length highly lower than in bulk technology [20, 21].

As the complete electron channel lies in the thin volume between the gate and the BOX, the gate voltage better controls the channel evolution. Thanks to that, the body effect $n$ and many short channel effects, as the DIBL, are dramatically reduced, thus providing an enhanced subthreshold slope [22, 23].

As there is no more PN junction between the bulk and the source or the drain of the FDSOI transistor, the absolute value between the source voltage and the back plane voltage, or back bias, can be superior to 0.3V. Moreover, it has been proved that the buried oxide can support a voltage difference of up to 2 volts without electrical breakdown [24]. However, the back interface of

FDSOI transistor is thin enough in such a way that a variation of the back bias influences the electrical characteristics and performances of the transistor. As with conventional bulk transistor, the gate voltage needed to form the electron channel depends on the voltage difference with the bulk voltage. In other words, the back bias modifies the threshold voltage of transistor. All of that means that, during the life time of the circuit, the threshold voltage of the transistor, one of its main parameter for the current-voltage dependency, can dynamically vary over a relative wide range. As the buried oxide is approximately ten times thicker than the main gate, the back biasing can be considered as a gate voltage having an influence roughly ten times lower on the current.

On the contrary of bulk transistor, the FDSOI transistor channel is undoped. It provides a lower variability - up to two times smaller $A_{V_{th}}$ [25] - because the threshold voltage does not depend any more on the dopant number (see section 1.2.2). As thoroughly explained in [26], a multi-$V_{th}$ technology is provided by changing the type of doping of the back plane and its polarization. Therefore, it is possible to produce up to three different threshold voltages without doping the channel and adding variability, unlike the bulk transistors. As a example, in the LVT feature of the 28nm FDSOI technology, the NMOS transistor is lying above a N doped well and the PMOS is above a P doped well. This feature will be used in every simulation and measurement results of this book.

Let us sum up the advantages of FDSOI technology owing to the addition of the BOX and the better channel electrostatic control:

- The parasitic junction capacitances are radically reduced.
- The subthreshold slope is higher, so is the $I_{on}/I_{off}$ ratio.
- There is no channel doping leading to a dramatically lowered variability.
- The threshold voltage can be modified by the back plane doping type and also dynamically by the back plane voltage (back bias) over a wide voltage range.

In [1], authors show that this technology offers several advantages and degrees of freedom to reach a pareto-like optimum in the energy-delay domain. These conclusions on the CMOS standard cells obviously apply for the most critical one in term of impact on the overall circuit performances: the flip-flop. In this respect, the next section will present the state of the art flip-flop architectures and will select the architectures which seem to be the most interesting ones for high-speed and low-power applications.

## 1.3 FLIP-FLOPS IN MICROPROCESSOR ARCHITECTURE

Flip-flop is the fundamental element of the synchronous logic. In a synchronous microprocessor, the clock signal is spread among a big number of paths all around the chip, forming the clock tree. The clock signal essentially brings in its edge the information that the data is valid and ready, and synchronises the whole circuit.

This section first explains why the number of flip-flops has literally exploded in the modern synchronous digital circuits and why their characteristics have a direct and huge impact on the overall circuits performances. Afterwards, the FF main figures of merits will be defined and then compared between the different flip-flop architectures of the state of the art. Thanks to that, the choice of explicit pulse-triggered flip-flop is motivated for high speed and low power applications.

### 1.3.1 Flip-flop contribution in modern ICs

The technique of pipeline has been widely and efficiently used to increase both the clock frequency and the operations per second of a processing unit, thus the instruction level parallelism (ILP). In contrast to the single instruction per cycle architecture, pipeline technique adds register along the datapath in order to separate the instruction treatment into several stages: fetch, evaluate, memory write-back,... It allows to reduce the length of the datapath between flip-flops and to treat different instructions at the same time. The clock frequency and the average throughput are thus intensively increased. Nevertheless, each pipeline stage needs a register to store all the data and the instruction control bits.

In addition to pipelined CPUs, the super-scalar technique has also been developed to increase the ILP. It consists of replicating combinational elements, the arithmetic/logical unit for example, to process several data in parallel. As the intermediate results of the ALU must be stored, the register file is also expanded in relation with the hardware replication. The choice in architecture has a direct and immediate impact on the number of flip-flops in the core: if the number of stages and/or of hardware replication is N times bigger, the number of flip-flops is roughly N times bigger.

Over the last years, systems proposing the *time-borrowing* technique to push away the speed limit of digital circuits have appeared in the literature. The idea of the time-borrowing technique is to work with a lower clock period than the critical path delay while sensing the end of critical paths thanks to custom flip-flops. If the circuit senses that a data changes during the sensing time after the triggering clock edge, a recovery mechanism is activated to guarantee the proper functionality of the circuit.

Therefore while increasing the speed, these three techniques induce a much higher complexity of the clock tree. Combined with the increase of the length of the datapath, tens of thousands of flip-flops were added to digital circuits to meet the speed target. Consequently, the energy consumption of the clock tree including its leafs (FF) exploded to become the most important part of the

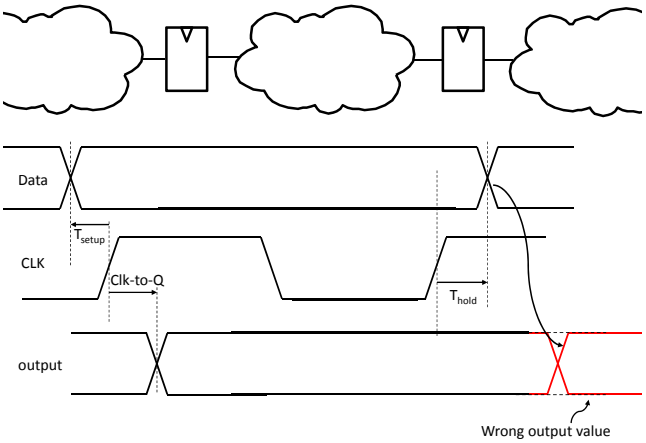



**Figure 1.5:** The flip-flop principle and its figures of merit.

overall energy consumption of a modern digital circuit, up to 50% or even 70% [4, 5]. On top of that, around 80% of the switching energy of a clock tree is located at the leaf level: the flip-flops [27]. The flip-flop architecture is thus an important element that designers must take into account to provide an energy-efficient circuit.

### 1.3.2 Flip-flop figures of merit

For standard high speed and low power applications, the FF figures of merit are the following (illustrated in figure 1.5):

- The clock-to-output (Clk-to-Q) delay. It is the propagation delay between the triggering clock edge and the time when the output data of the flip-flop is valid. It should obviously be as short as possible.

- The setup time. The setup time is the minimum time when the data must be valid before the triggering clock edge. It is defined *positive* when the data edge is *before* the clock edge and should be as low as possible. Designers use metrics based on the three signals of the figure 1.5, to provide quantitative criteria to compute the setup time. We propose a new metric dedicated to pulse-triggered flip-flops in [28].

- The input-to-output (D-to-Q) delay. For flip-flops, it is the time between the input data edge when it arrives at the setup time, and the output data edge. Therefore, the D-to-Q delay can be written as equal to the Clk-to-Q delay plus the setup time (D-to-Q = Clk-to-Q + setup).

- The hold time. The hold time is the minimum time until when the valid data must stay valid and stable after the triggering clock edge. It is defined *positive* when the data edge is *after* the clock edge.
- Finally, the energy consumption and the area of the flip-flop should obviously be as low as possible.

### 1.3.3 Flip-flop architectures

There is an extremely large amount of flip-flops topologies in the literature. Edge-triggered cells have been the subject of many studies and continue to feed the imagination of researchers.

In [6], a comparison of nineteen state-of-the-art flip-flops is performed. The authors first classified them into four main categories:

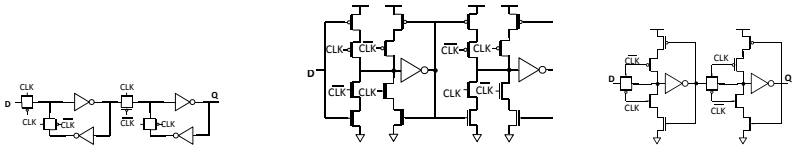
- master-slave flip-flops
- differential flip-flops
- pulse-triggered flips-flops: implicit and explicit
- dual-edge triggered flip-flops

This section shows, for each topology, the different architectures proposed in the literature and concludes by a summary of their advantages and drawbacks. Then, the next section explains more precisely why we selected the explicit pulse-triggered flip-flop architecture category as the best candidate for our high-speed and low-power applications.

#### *Master-Slave*

The master-slave flip-flops are composed of two latches connected in series. The input data is connected to the master latch and the output of the FF is the output of the slave latch. The master latch is enabled during the clock period preceding the triggering clock edge, so that the data information can pass through the latch during this period. The slave latch is enabled during the other period of the clock. Consequently, the information passes from the master latch to the slave latch only at the triggering clock edge. During the non-triggering clock edge, the slave latch becomes closed and keeps the previous data value while the master latch becomes open and can switch to a new data value.

As soon as the master latch samples the correct data, the slave latch will automatically follow. Therefore, the setup time is highly subject to the switching speed of the master latch. Then, the clock to output (Clk-to-Q) delay depends mainly on the propagation delay in the slave latch. The input-to-output (D-to-Q) delay, here the sum of the setup time and the Clk-to-Q delay, is not only the propagation delay across the two latches but has also to take into account the margin (included in the setup time) needed to guarantee the correct latching of the data.




**(a)** The original and conventional transmission gate master-slave (TGMS) architecture.

**(b)** Complementary CMOS ($C^2MOS$) architecture. Pass-through phenomena avoided.

**(c)** The conventional transmission gate flip-flop (TGFF) architecture.

**Figure 1.6:** Conventional master-slave architectures. TGFF gives the best tradeoff between speed performances and pass-through phenomena.

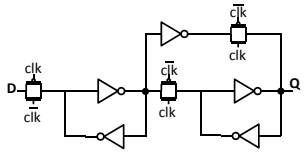
Among a large panel of master-slave FF in the literature, let us remind the conventional static architecture shown in figure 1.6a. This architecture is called the transmission gate master-slave (TGMS) and is the basic structure of master-slave flip-flop.

The great advantage of the complementary square MOS ($C^2MOS$) architecture, compared to the TGFF one, is its immunity to the clock overlapping phenomena. Indeed, if the complementary clocked signals ($clk$ and $\overline{clk}$ on figure 1.6) are temporarily both at the same level, there is a short circuit path between the inverter supplies and the input data of TGMS whereas it is impossible for the $C^2MOS$ structure. This effect creates a short circuit current during a short period of time which can nevertheless represent a significant energy overhead.

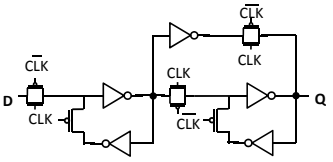
In order to combine the speed of the TGMS -thanks to pass-gate utilization- and the robustness of the $mC^2MOS$, the transmission gate flip-flop (TGFF) has been proposed and is today the reference for master-slave architecture comparison (figure 1.6c).

More complicated master-slave architectures have been proposed in the literature. We can cite the comparisons of [9], [29], and [30] and the proposition of [31], both illustrated in figure 1.7. The last two propose NAND-gates based master-slave flip-flops. Those architectures do not have any transparency time thus avoiding metastable state [9, 31, 32]. Nevertheless, the propagation delay is highly increased by the series of NAND gates and thereby these structures are not used in applications needing performances.

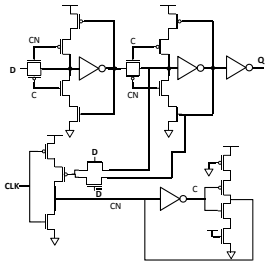
Seeing the principle of master-slave flip-flop, we can say that the setup time is in nominal conditions strictly positive, meaning that the correct data value must change before the clock edge in order to permit the master latch to sample the data before its closing. On the other hand, the hold time is, in nominal conditions, negative because once the master latch is closed after the triggering clock edge, a change of the input data will not influence the flip-flop output.



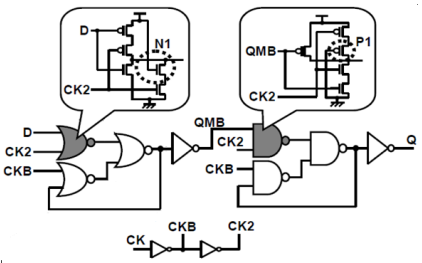

**(a)** The push-pull D-flip-flop [29].



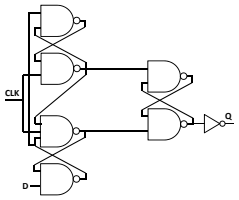

**(b)** The push-pull Isolation D-flip-flop [29].



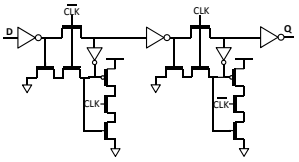

**(c)** The gated master slave FF (GMSL) [30].




**(d)** Contention-less flip-flops (CLFF) dedicated to specific applications [31].




**(e)** Race-Free NAND-based D flip-flop [32].




**(f)** The write-port master slave FF (WPMS) [33].

**Figure 1.7:** Novel master-slave architectures proposed in the literature.

### *Differential*

The differential flip-flops, also called dual-ended, have also the complementary input data as input and generate simultaneously the output signal and its complement. The primary concept of the design is the symmetry of the structures, ensuring a identical behaviour between the data-to-output path and the complementary path. This concept is illustrated in figure 1.8 by the adaptation of the conventional master-slave structure.

In non differential architectures, an extra inverter might be used at the synthesis step if the complementary output is needed in a following path. This causes a speed penalty and a shift between the two complementary signals, possibly

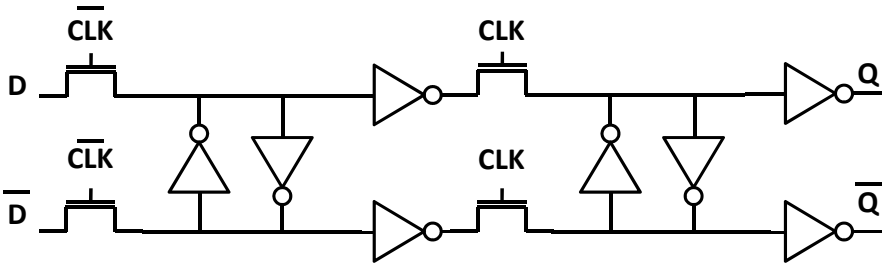



**Figure 1.8:** The differential flip-flop principle (master-slave-based).

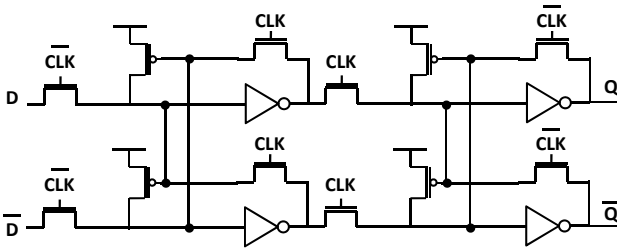
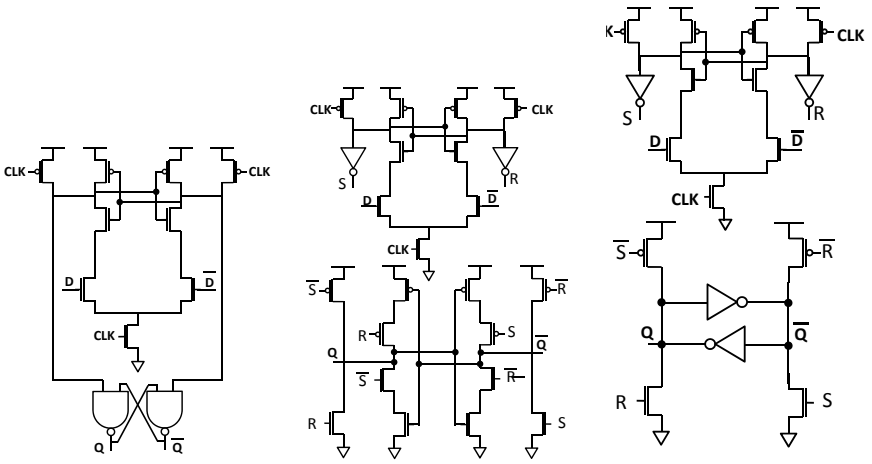



**Figure 1.9:** Improved MS-based differential structure [34].

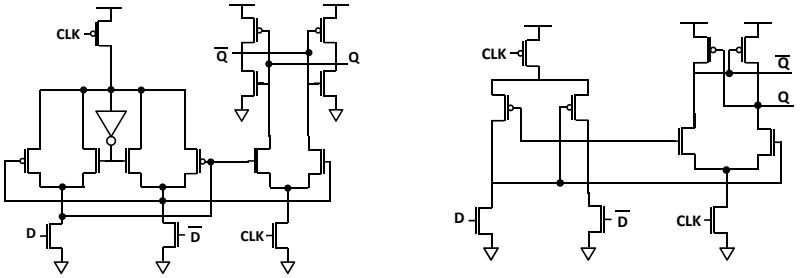
causing disturbing glitches later in the circuit [34]. That is why the main application of this kind of flip-flop is the deeply pipelined systems where the delay from flip-flop is a heavier penalty for the computational time [34]. Moreover, the improved alignment in the output signals is especially suitable for combinational elements such as decoders and multiplexers, where the complementary signal is used in every computation.

Despite some exceptions (figure 1.9), most of the differential flip-flop found in the literature are based on the sense amplifier principle (figures 1.10): one differential pair is connected to a symmetrical bistable element and senses the complementary inputs. The nodes of the amplifier part are pre-charged to the supply voltage value ($V_{dd}$) during the clock level preceding the triggering-clock edge. After that, the high input value pulls one dynamic node to ground which switches the following bistable to its new state. During the following clock level, the flip-flop is always open and can evaluate a new data coming to the input. It means that this kind of flip-flops are actually latches and designers must take care of the hold time constraint and the glitch effects. After the non-triggered clock edge, the dynamic nodes are pre-loaded again.




**(a)** SAFF basic structure [35]. **(b)** Improved SAFF topology [36]. **(c)** With modified output part.

**Figure 1.10:** Differential sense-amplifier based flip-flops.




**(a)** Static single-transistor clocked (SSTC). **(b)** Dual single-transistor clocked (DSTC).

**Figure 1.11:** Differential master-slave flip-flops [35].

To avoid the level-triggered behaviour of the previous architecture, [35] proposed differential master-slave architectures (figure 1.11). Nevertheless, those structures suffer from a highly penalizing voltage drop at one of the slave latch input, thus reducing the driving capability of its outputs [8].

Thanks to the NMOS differential pair, differential flip-flop architectures have shown lower D-to-Q delays than master-slave ones. Nevertheless, because of

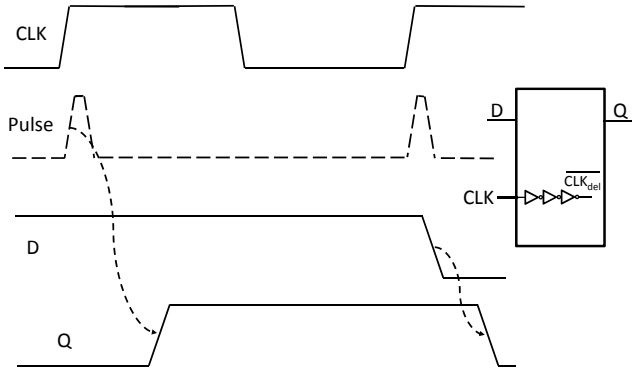



**Figure 1.12:** The pulse-triggered flip-flop principle. During the pulse period, the flip-flop behaves essentially like a level-sensitive latch.

the pre-load, the additional complementary data-to-output path and the delay buffers needed to fix the hold time constraint, this topology leads to a huge energy overhead compared to master-slave. Furthermore, in many differential architectures, there is no feedback transistors in the bistable element. It means that a stack of NMOS must fight and beat a PMOS in the bistable element in order to change its state. Consequently, those structures are extremely susceptible to process variation at ultra-low voltage where an unbalanced strength may cause serious issues and design problems. For these reasons, differential flip-flops are not suitable for low-power applications and have essentially been used in high-speed processors.

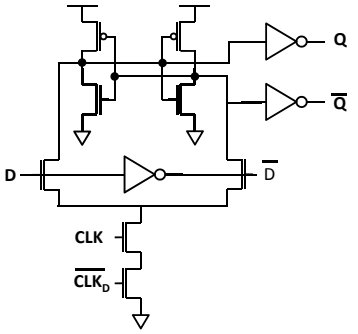
### *Pulse-Triggered*

The pulse-triggered flip-flops, or simply pulsed-flip-flops (pulsed-FFs), are composed of only one latch which is open during a short period of time after the triggering clock edge, defined by a pulse-like signal. During this short period, the data value passes through and changes the state of the latch, as illustrated in 1.12. Afterwards, the latch is closed and its output value remains constant.

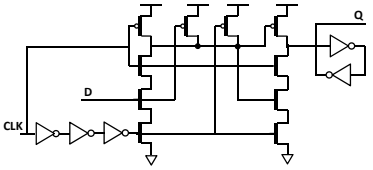
The pulsed behaviour is achieved thanks to a delayed clock signal generated in the flip-flop. During the period defined by this delay, both the clock and the delayed clock signals have the same voltage value. A comparative stack allows to get open the latch during this period.

Due to the architecture, a data value changing slightly after the clock edge can be sampled by the latch. Therefore, the setup time is essentially negative while the hold time becomes positive in contrast to the master-slave topology. As only one latch is laying in the D-to-Q path and no timing barrier is involved, the speed of pulse-triggered FF is largely superior to MS.

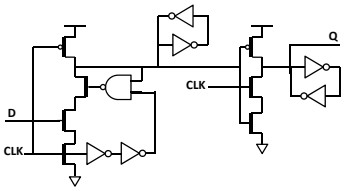
There are two main types of pulsed-flip-flops: implicit and explicit.



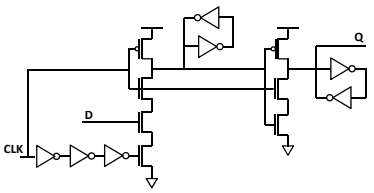

**(a)** Sense-amplifier based implicit pulse-triggering structures [37].




**(b)** Hybrid-latch flip-flop (HLFF) [8].




**(c)** Semi dynamic flip-flop (SDFF) [8].

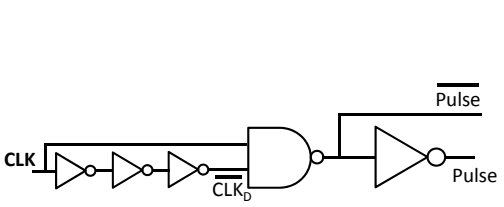



**(d)** Implicit-pulsed data-close-to-output (ipDCO) semi-dynamic hybrid flip-flop [7].

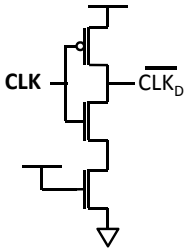
**Figure 1.13:** Differential and semi-dynamic implicit pulse-triggered flip-flops.

In implicit pulsed-FFs, there is no node which exhibits a pulse signal at any time during the clock cycle (see figure 1.13). The pulse behaviour is thus called implicit. As the charge and discharge into the flip-flop also depend on the data, the architectures often need a stack of minimum three transistors: two performing the pulse property and one containing the data information. A lot of proposed implicit pulse-triggered flip-flops have been called semi-dynamic, as reference to the dynamic NMOS logic. Those structures are composed of a static part, meaning a bistable element, and a dynamic part with a NMOS stack and a pre-charge mechanism (figures 1.13b, 1.13c, and 1.13d).

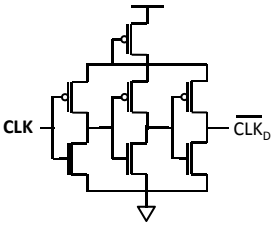
For the explicit pulsed-FFs, the latch is connected to the output(s) of a pulse generator (PG). The element provides at least one signal which has effectively the shape of a pulse. It often contains the delay generator and has the advantage to be shared by several latches. As the latches need only the pulse signal to perform their flip-flop functionality, like the clock signal for master-slave flip-flop, this signal can be propagated as well as the clock tree in a conventional digital circuit. Moreover, a stack of two transistors in the latch is sufficient to contain



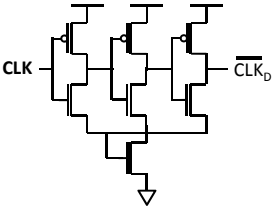

**(a)** The conventional pulse generator.




**(b)** Low area delay generator [38].




**(c)** Low swing PMOS version.

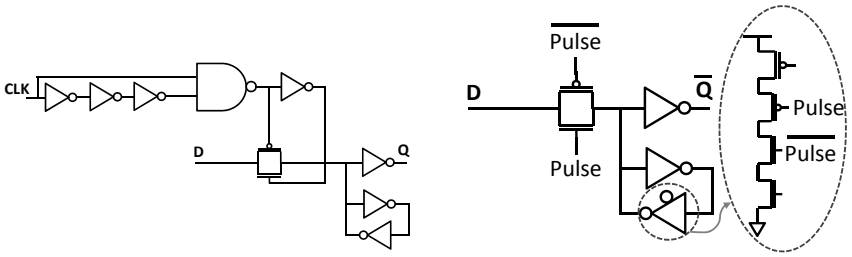



**(d)** Low swing PMOS version.

**Figure 1.14:** Pulse generators dedicated for pulsed-FFs in the state of the art. In 1.14c and 1.14d, transistors in diode mode lead to a lower swing, thus energy consumption, in the delay generator [39].

the clock and input data informations. The overview starts by the different pulse generators found in the literature and then the latches. These latches can be used with every presented PG and conversely.

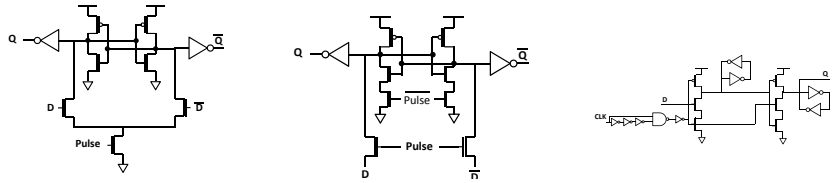
The conventional pulse generator for pulse-triggered flip-flops is shown in figure 1.14a. The inputs of the NAND gate are the clock signal $CLK$ and its delayed complementary signal $\overline{CLK_d}$. It provides the complementary pulse signal $\overline{Pulse}$ which is connected to the latch and to the input of an inverter providing the $Pulse$ signal. The delay generator is located between the two inputs of the NAND and is composed of three inverters giving the complementary and delayed signal needed. In [38], authors propose to add stacking in the delay path in order to use only one inverter to perform the delay. It has also the advantage not to increase the clock input capacitance (clock load) on the contrary of increasing the gate length. The pulse generators of figures 1.14c and 1.14d have been implemented in an implicit pulsed-FF [39] but this low swing technique can be easily transposed in the delay generator of the conventional PG.




**(a)** The simplest expression of a explicit pulse-triggered flip-flop [7].

**(b)** The conventional explicit pulse-triggered flip-flop (ep-FF) [40].

**Figure 1.15:** The basic structures of explicit pulse-triggered flip-flop.




**(a)** Sense-amplifier based explicit pulsed-FF [41]. Here, the stack is only two NMOS.

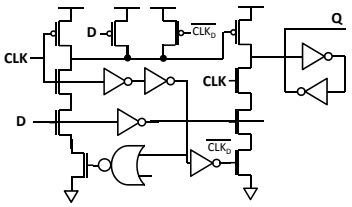
**(b)** The feedback of the bistable is here weakened [40]

**(c)** Basic semi-dynamic explicit-pulsed flip-flop [7]

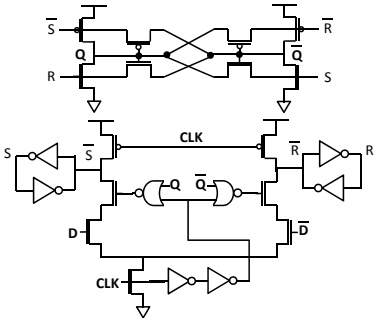
**Figure 1.16:** Differential and semi-dynamic architectures of explicit pulsed-FF.

The extremely basic explicit pulse-triggered structure is shown in figure 1.15a. A great improvement in its robustness is shown in figure 1.15b where the feedback of the bistable element is idle during the pulse. Those structures are also called in the literature transmission gate pulsed latch (TGPL) to make the differentiation with the master-slave structure TGFF (figure 1.6). As with implicit pulsed-FFs, sense-amplifier-like versions have appeared in the literature. As designers always inspire from previous topology, a semi-dynamic explicit pulsed-FF has also been proposed ([7] figure 1.16c).

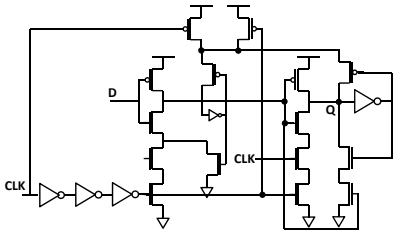
*Conditional techniques* In many architectures seen before, especially in the semi-dynamic topologies, nodes are charged or discharged even when the data does not change. As the data switching activity factor ($\alpha_{sw}$) is on average between 0.1 and 0.4 depending on the application, this leads to a useless energy consumption. Therefore, the interesting idea of conditional latching technique appeared in the literature. The key point is to use the information of the output of the flip-flop, combine it with the input data and enable or not the switching, to save the



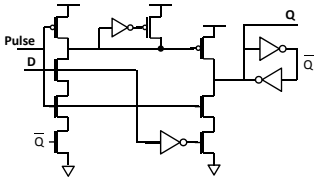

**(a)** The singled-ended version of conditional capture flip-flop (CCFF) [42].




**(b)** Dual-ended or differential version [42].




**(c)** Conditional precharge flip-flop (CPFF) [43]




**(d)** Conditional discharge flip-flop [44]

**Figure 1.17:** Other conditional techniques.

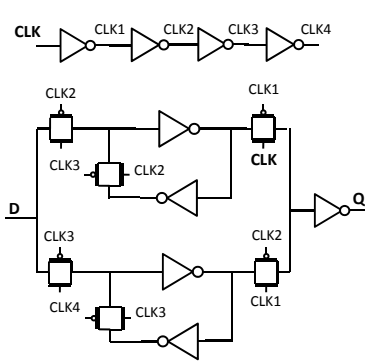
energy consumption. As said, the main characteristic is the use of the output signal in feedback in the writing system.

In semi-dynamic structures, we can cite the conditional capture technique represented in figures 1.17a and 1.17b and the conditional precharge and its complementary discharge techniques in figures 1.17c and 1.17d. Each of them was implemented in pulse-triggered topology.

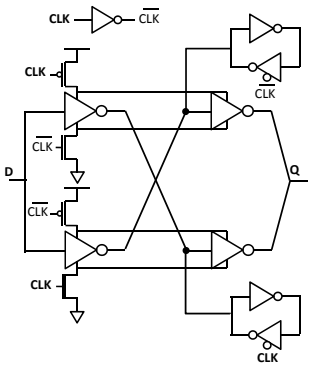
### *Dual-edge*

A dual-edge flip-flop latches the data on both edges of the clock. With the same datapath, the clock cycle can be divided by two, keeping the same throughput. As the switching activity of the clock tree is divided by two, so is its switching energy.

Every architecture discussed in the preceding section can be adapted in dual-edge-triggered topology [7, 45]. For most of them, the writing part is doubled to perform the triggering on both edges, meaning that the internal energy con-

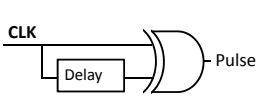



**(a)** With this implementation, dual-edge version leads to a higher overall energy consumption.

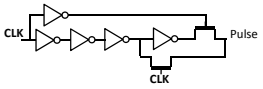



**(b)** The conventional latch duplicated.

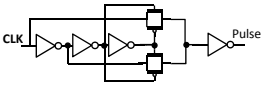
**Figure 1.18:** A basic implementation of the dual-edge property [7].




**(a)** The basic dual-edge pulse generator concept.




**(b)** Dual-edge PG with NMOS-only pass-gates [41].




**(c)** Dual-edge pass-gated based pulse generator [46].

**Figure 1.19:** Dual-edge pulse generators.

sumption of the flip-flop is increased. At some point, the data and clock loads are so increased and even doubled that the saving energy in the clock tree does not cancel out the energy overhead due to making the dual-edge feature (see [7] and figure 1.18a).

For all the reasons cited above, the literature trend is the adaptation of explicit pulsed-FF into dual-edge version. The main change in the architecture of dual-edge flip-flops is the pulse generator, working on both edges. The direct solution is to put a XOR gate instead of an AND (NAND+inverter, figure 1.19a). Other implementations are shown in figure 1.19.

Finally, let us notice that a duel-edge triggering system must ensure a perfect balanced clock cycle. Indeed, it is the shorter level of the clock which determines the operational clock frequency. Therefore, buffers and inverters must have perfectly balanced rising and falling propagation delay, as well as the clock generator. This assumption is extremely hard to guarantee because of process, voltage and temperature (PVT) variations, especially at ultra-low voltage.

*Summary and state of the art in ultra-low power domain*

The master-slave architecture is the most used topology, especially in low power circuits. Its behaviour is robust and straightforward to understand for micro-architecture designers and that is why it is today the most used FF architecture, especially in ultra-low voltage circuits. Nevertheless, it suffers from a large D-to-Q delay because of the use of two latches virtually separated by a clock barrier.

The differential flip-flops are composed of two symmetric rails of data and are based on the sense-amplifier principle. They are more susceptible to process variation and often need a precharge at each clock cycle exhibiting a high energy consumption. They are so highly not recommended for ULP and ULV circuits.

The pulse-triggered flip-flops are made of only one latch which is open during short period after the triggering clock edge. The propagation delay is thus dramatically reduced and the setup time becomes negative while the hold time is positive.

A dual-edge architecture samples the data at each edge of the clock signal. Some architectures allow to reach this property without duplicating the datapath and thus doubling the clock and data load. But the most energy-efficient way is to modify the pulse generator of the explicit pulsed-FFs. However, the clock cycle must be perfectly balanced in order to have the same timing constraints both in high and low levels of the clock. This statement is hardly achieved at ultra-low voltage.

Table 1.1 gives a summarized comparison of the four flip-flops topologies.

**Table 1.1:** Comparison of the flip-flop topologies for different figures of merit.

| Topology | Speed | Power | Area | $V_{dd}$ Scalability |
|---|---|---|---|---|
| Master-slave | – | + | + | ++ |
| Differential | ++ | – – | – | – – |
| Implicit pulsed-FF | + | – | – | – |
| Explicit pulsed-FF† | ++/++ | -/++ | -/++ | -/- |
| Dual-edge | + | ++ | + | – |

† without/with shared pulse generator

Let us now highlight the state of the art about the architectures reaching good performances in low-power operations.

First, we can cite the numerous conditional techniques presented before [42, 43, 46, 47], helping to decrease the energy consumption. Then, in the wide comparison performed in [6], the basic and conventional transmission-gate structures (see the TGFF in figures 1.6c) are pointed out to be the most energy-efficient ones.

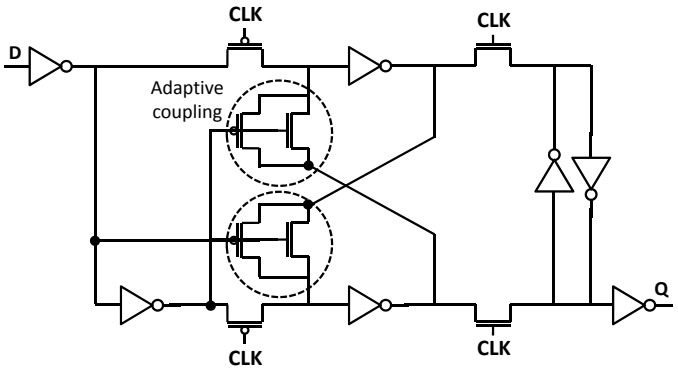



**Figure 1.20:** Adaptive-coupling flip-flop [48]. Thanks to the elementary clock load, this architecture exhibits the lowest energy consumption among the flip-flop literature.

Later, the adaptive coupling flip-flop (ACFF) was proposed by [48] to tackle the susceptibility of differential structures to process variation. As we can see in figure 1.20, the basic differential structure is enhanced by the addition of two pairs composed of one NMOS and one PMOS in parallel. This helps to alleviate the drawbacks of the conventional topology, due to a short circuit path formed between supplies and an overdesign of the pull-down NMOS for worst case conditions. As the conventional structure, it needs only one phase of the clock which means the saving of clock buffer and two times less clocked transistors, meaning reduced clock load. This architecture has been fabricated in 40nm CMOS bulk process and has shown an energy per cycle of 2fJ per cycle for a data activity of 10% at 1.1V which is the lowest measured energy consumption reported in the literature at nominal voltage. The main drawback of this structure is its high setup time leading to a relatively large D-to-Q delay (see table 1.2). Finally, silicon measurements showed a tolerable yield down to 0.6V.

Later on, [49] proposed the conditional push-pull pulsed latch ($CP^3L$) architectures and studied it at nominal voltage operations (1V). It can be categorized as an explicit conditional pulse-triggered flip-flop with the output feedback in the pulse generator (see figure 1.21). It exhibits the best energy-delay product (EDP) among the literature, mainly thanks to the small delay of pulse-triggered flip-flop (table 1.2). As the D-to-Q path is composed of two stages, it exhibits better performances for a high output capacitance [49].

[50] proposed a differential flip-flop (IMD-FF) and studied its reliability until 0.4V. Unfortunately, it is only compared to other differential flip-flops and exhibits a NMOS stack having to counter a PMOS.

Finally, [51] proposed an adaptive pulse-triggered flip-flop (APFF) with a replica delay generator with the aim of performing operation over a wide voltage

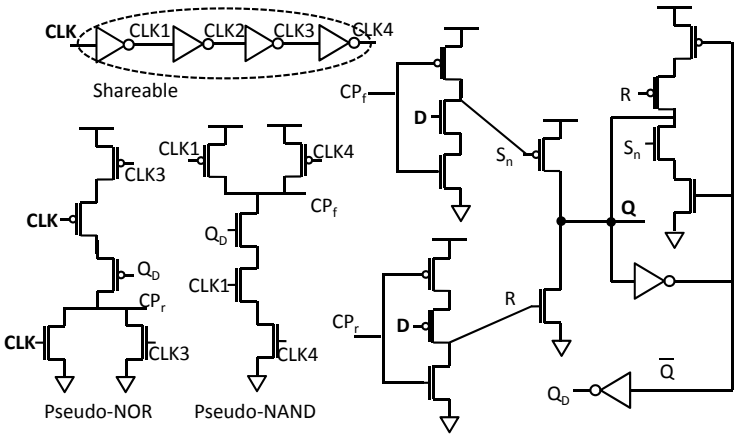



**Figure 1.21:** Conditional push-pull pulsed latch ($CP^3L$) [49]. The lowest EDP of the literature.

range, from 1V down to 0.2V. Comparison to conventional master-slave has shown that, over the wide range, the delay is more than two times lower, the energy consumption about 30% higher, and thus the energy-delay product of the pulsed-FF represents approximately 65% of the master-slave EDP.

**Table 1.2:** A sample of the quantitative results of the state of the art flip-flop performances

| Architecture | Supply [V] | Technology | D-to-Q [ns] | Energy [fJ] [data activity] | EDP [fJ.ps] | Transistor count |
|---|---|---|---|---|---|---|
| ACFF* | 1.1 | 40nm | 0.264 | 2 [10%] | 528 | 22 |
| $CP^3L$* | 1 | 65nm | 0.017 | 26.1 [10%] | 451.5 | 34 |
| IMD-FF | 0.4 | 90nm | 1.7 | [n.a.] | [n.a.] | 22 |
| APFF | 0.4→1V | 65nm | 14.2 | 4.2 [n.a.] | 59700 | 30 |

* silicon measurements

## 1.4 THE EXPLICIT PULSE-TRIGGERED FLIP-FLOPS

Keeping in mind all of the descriptions above, this section points out the remarkable properties of the explicit pulse-triggered flip-flops and motivates this choice for the research strategy of the following chapters.

As every pulsed-FF, there is only one latch in the input-to-output path. The propagation delay is thus extremely low compared to architectures with two complementary latches.

In opposition to differential flip-flops, only one data rail and output inverter are charged and discharged. Moreover, contrary to dynamic or NMOS-like flip-flops, there is no pre-charging. The energy consumption is so largely lower than the second category of FF.

Compared to implicit pulse-triggered flip-flops, the explicit version should provide a lower delay for the same sizing because of the additional stack needed by implicit topology. This stack is also a drawback for ULV operations as it decreases the robustness of the CMOS logic. On the other hand, the pulse generator needed by explicit structure should provide a higher energy consumption with the same transistor size. Nevertheless, the output(s) of the pulse generator could be sent to the neighbouring identical latches, *i.e.* flip-flops. This technique provides a reduction of both the energy per flip-flop and the area per FF because the consumption and area of the pulse generator is normalized by the number of shared latches.

The dual-edge version of explicit pulse-triggered flip-flop only needs a dual-edge version of the pulse generator, as the latch only needs a pulse signal at each clock edge to properly work.

Finally, explicit pulsing easily allows to perform time borrowing technique. In this text, time-borrowing has no concern with the latch-based pipeline architectures. It means that the valid data might arrive during a time-borrowing window after the triggering clock edge. If a data transition is detected during this time-borrowing window, an error signal is sent to the controller of the microprocessor architecture. Then, an error detection method with or without an error correction mechanism, is used to handle the late data arrival. The figure 1.22 shows a direct implementation of the time-borrowing and clock stretching techniques developed at system level in [52] [53] and [54]. Here we use one pulse generator, one latch and one transition detector which needs less hardware than one master-slave flip-flop (2 latches), one latch and one XOR gate as proposed in [52] and [54]. Furthermore, the latch connected to the pulse generator has a flip-flop behaviour. So, the delay error detection window is the pulse width and not nearly the half of the whole clock cycle as in [53]. This property avoids a massive additional buffering insertion as mentioned in [53]. Change needed in the pulse generator simply consists in the addition of two minimum sized transistors to perform the NAND gate behaviour. A control signal maintains the pulse active during the desirable time such as the next latch in the datapath stays open. This control signal is generated in the transition detector or somewhere

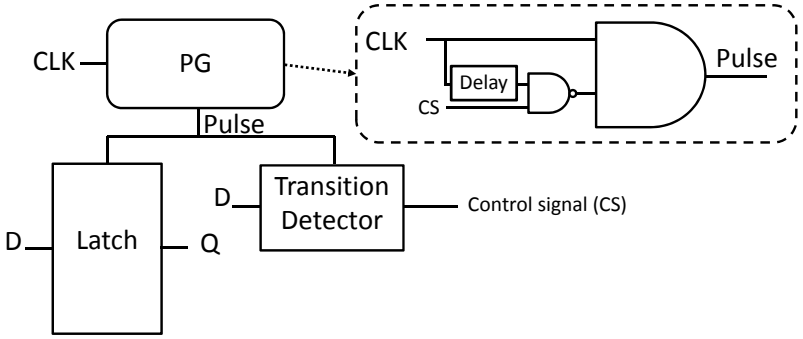



**Figure 1.22:** Example of implementation of the time-borrowing technique in explicit pulse-triggered flip-flop.

else in the circuits, and provides a clock stretching property without the need of synchronous signals as in [54].

Nevertheless, let us not forget that the pulse-triggered topology exhibits several disadvantages.

As the pulse signal makes the latch open after the triggering clock edge, the hold time is essentially positive, contrary to master-slave FF. Indeed if the data value quickly changes after the clock edge, a wrong data will be latched. Then, during the synthesis placement, the tools will automatically add buffer in short path to fix the hold constraints. This drawback can be mitigated by the simultaneous negative setup time thanks to the useful skew technique [55, 56] and delay buffers designed with the current-starved technique (Chapter 3).

Then, it directly follows from the architecture principle that, if the pulse width is not wide enough, the latch will not have enough time to switch with the new data. The issue is especially critical in the ultra-low voltage domain where the local variations lead to a wide range in the pulse width. This problem will be fully discussed and treated in Chapter 3.

## 1.5 CONCLUSION

In the aim of ULP and UWVR circuits, the need of efficient clock tree is of primary importance. This chapter has presented the FDSOI technology and the state of the art of flip-flop topologies and motivated the choice of explicit pulse-triggered flip-flops in FDSOI technology for the targeted ULP and UWVR applications.

In order to maintain the trend of the microelectronics industry, the new FDSOI technology has been proposed to overcome the bulk limitations. This technology allows designers better electrical performances, lower variability and a powerful degree of freedom - the back bias voltage - which allows to change

dynamically the threshold voltage of transistors. Several previous studies have demonstrated the benefit of this technology compared to the conventional bulk one, especially to provide energy-efficient circuits.

In a modern complex digital circuit, the number of flip-flop and the complexity has literally exploded due to the increasing number of bits in the datapath and the intensive use of pipeline and super-scalar techniques. The proper flip-flop architecture is thus a challenge for designers who want to reach a good energy consumption under throughput constraints. Among a wide overview of the flip-flop topologies, the explicit pulse-triggered architecture was pointed out as an extremely interesting candidates for high-speed and low-power applications, thanks to:

- a very small input-to-output delay - defined when the input arrives at the setup time,
- a negative setup time allowing much more reduction of clock cycle,
- the sharing of the energy-consuming pulse generator, reducing both energy and area,
- a lower delay than implicit pulsed-FF thanks to lower stacking,
- dual-edge and time-borrowing facilities that can be more easily implemented with explicit than implicit pulsed-FFs.

In the following chapters, the design of energy-efficient explicit pulsed-FFs in FDSOI is studied. As seen, the study can be separated between the latch and the pulse generator which will therefore be the topics of the next two chapters.

**CHAPTER 2**

# STUDY AND COMPARISON OF LATCH ARCHITECTURES FOR UWVR PULSE-TRIGGERED FLIP-FLOPS IN 28NM FDSOI

**Abstract**

This chapter gives a complete comparison of explicit pulse-triggered flip-flops architectures in order to select the most energy-efficient ones, according to the specifications of UWVR and ULP applications: fast and energy-efficient at nominal voltage and extremely energy-efficient at low voltage. First of all, a list of design constraints is elaborated to take into account the particularities of the targeted applications. From that, the choice of the compared architectures is motivated thanks to a theoretical analysis of the writing network in a latch. The sizing methodology used for the comparison is introduced and discussed.

After all that, the results of the comparison are exposed in the energy-delay domain. At nominal voltage, the TGPL-Data and $C^2$MOS-Data architectures are shown to be the most energy-efficient in the low-power region while the TGPL-Clk structure exhibits the best energy-delay product in the high-speed region. At ultra-low voltage, $C^2$MOS-Data presents the lowest energy consumption in every region excepted for very high speed applications. In order to deal with the fact that the results are not the same at nominal voltage and at ultra-low voltage, it is shown that the back biasing technique allows to dynamically reach both high-speed and low-power properties with the same architecture sizing. It means that the FDSOI technology allows a higher energy-efficiency than obtained by the sizing methodology thanks to the back biasing technique.

Next, we mention that none additional problem appears during the implementation in conventional flow, compared to hard-edge triggered flip-flops. Finally, silicon measurements of previously selected architectures in 28nm FDSOI highlight and confirm the previous conclusions. In particular, an average clock-to-output delay of 31ps has been measured for the TGPL-Clk architecture in nominal conditions.

**Contents**

## 2.1 INTRODUCTION

As seen in the previous chapter, among all the flip-flop topologies found in the literature, the explicit pulse-triggered architecture is the most promising candidate for the high-speed and low-power applications. As fundamental principle, the clock signal is connected to a pulse generator (PG) providing a pulse signal to a level-sensitive latch. The latch architecture is of fundamental importance in the flip-flop performances. Indeed, it directly gives the speed of the FF and directly impacts the sizing of the pulse generator, thus the overall energy consumption. In order to reach an energy-efficient FF, we have to find a latch providing the required speed for a minimum amount of energy.

In this chapter, a large set of pulse-sensitive latches is compared. First, Section 2.2 exposes the design constraints of the latches for this work focusing on ULV and UWVR circuits. These constraints come from the targeted ultra-low voltage applications, the characterization method used in ASIC design, and from the use of an aggressive technology. Based on the design constraints, Section 2.2 develops and presents the choices of latches that will be compared. It is motivated by a theoretical analysis on the writing system of pulse-triggered flip-flops (pulsed-FFs). In Section 2.3, the sizing methodology of the latches is firstly presented. It is shown that the optimum ratio between the width of the NMOS and PMOS transistors ($\beta_{PN}$) is not the same if we target identical timing or maximized robustness, while the CMOS stage stack slightly varies the conclusion. Secondly, the complete comparison is performed at nominal and ultra-low voltages and over the whole energy-delay domain. The results of this comparison are then discussed for each supply voltage. As the results are not exactly the same at nominal and ultra-low voltages, we provide a discussion about the choice and the desired tradeoff depending on the application. Consequently, the use of a wide back biasing range possible in FDSOI technology is added to the discussion. It is shown that it allows a better tradeoff in speed and energy than by sizing methodology. The final Section 2.4 shows silicon measurements of these selected architectures which confirms the previous finding.

## 2.2 SELECTION OF ULTRA-LOW VOLTAGE LATCHES

As we saw in Section 1.3, flip-flop architectures are extremely varied and can be optimal for a specific application. Therefore, we have to determine some choices and design constraints in order to select relevant latches for the targeted applications of this work. After describing and motivating these constraints, the elaboration of the writing system is analysed in Section 2.2. As we will see, the conclusions lead to pulse-triggered flip-flops already presented in the previous chapter. Slight improvements in some architectures are finally proposed.

### 2.2.1 Design constraints

First of all, the target of energy-efficiency leads us to choose the static complementary MOS logic (CMOS). This logic style has been present for three decades in the industry because it provides a strong immunity to crosstalk and a lower total energy consumption thanks to the absence of short circuit path [12].

#### *2.2.1.1 Ultra-low voltage operation*

The pulse-triggered flip-flops of this work have to function at ultra-low voltage (ULV) in order to provide a high energy-efficiency. At this operational mode, the on-state current is small, meaning low operational clock frequency, and varies dramatically with the environmental variations. Moreover, the leakage current becomes predominant and is even more strongly proned to environmental variations. All this leads to several design consequences.

A short circuit path during the writing time is forbidden. The process variations may lead to a strength difference between PMOS and NMOS extremely high compared to the nominal case. Therefore, designers have to increase the transistor dimensions to ensure an acceptable signal to noise margin in writing mode (SNMW), leading to an unacceptable energy overhead.

A pass-gate with only one transistor is forbidden. In some architectures working at nominal voltage, a NMOS (or a PMOS) performs both the charge and the discharge of the next node. Because of the configuration, the charge (discharge) is not complete and the node value reaches quickly $V_{dd} - V_{th,n}$ ($gnd + V_{th,p}$) but the rest of the transition becomes extremely slow. For sufficiently high voltage, this value is high (low) enough compared to the threshold voltage of the next CMOS branch. But, at ultra-low voltage, the transistors would pass quickly in cut-off mode and the time needed to reach the switching threshold voltage of the next branch might be several orders of magnitude higher than standard operations.

A stack of four transistors is also forbidden. Stacking transistors decreases the speed and the robustness of the gates if the size of the transistors is not increased. Therefore, designers have to make a tradeoff between the speed-robustness and the area of the standard cells. We chose in this work a maximum stack of three as it has already been the case in other works in ULP operations [57].

Finally, the flip-flops of this work will not be dual-edge-triggered. In dual-edge systems, the clock tree must be perfectly balanced, providing an equal time in the low level and the high level of the clock. Nevertheless, if the constraint of 50%-50% is relaxed, it will always be possible with explicit pulsed-FFs to adapt the latch in dual-edge mode.

#### *2.2.1.2 Logic synthesis aspects*

Even if our flip-flop comparison is a study at gate level, the objective of every digital cell is to be implemented in a complete circuit. Consequently, we have to keep in mind that the flip-flops will be characterized and then used for synthesis

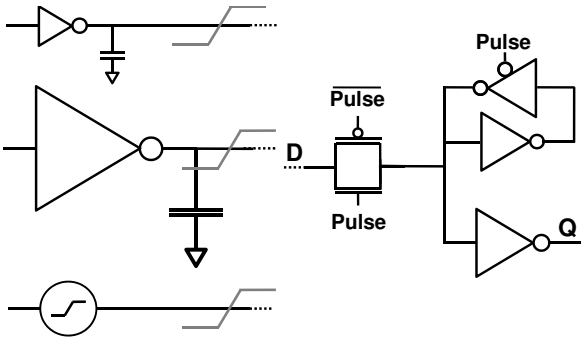



**Figure 2.1:** Three different configurations providing identical input slopes when the transmission gate is closed.

by CAD tools. These points lead us to set out two constraints which are not always found in the literature.

The output of the flip-flop must be the output of an inverter. Indeed, if the FF output is a node of the bistable element, the time needed to (dis)charge it, *i.e.* the minimum pulse time, depends on the output load. As the output load of a standard cell is only known after the synthesis placement, it might lead to two situations: (i) if the load is too large, the pulsed-FF could not be able to switch the state of its bistable element during the pulse period; (ii) if the load is too weak, the flip-flop will be overdesigned, meaning useless energy overhead. The size of this inverter will give the driving strength (or drive) of the cell and is an input of the sizing algorithm.

All of the external inputs arrive on a transistor gate. In a lot of latches proposed in the literature, the data input of the analysed flip-flops is the source/drain contact of a transistor. This directly means that there is a stack of two, at least, between the other source-drain contact of the transistor and the supplies; this might lead to several problems. First of all, is the input slope chosen when the transistor is on or off ? If it is on, it will not be the case in most of the characterization tool. If the clocked transistor is off during the input transition, the drive of the stack - when it will be on - will directly impact the time of charge or discharge and thus the timing performances of the flip-flop. But, for two identical input slopes, the stack may be very different as well as the corresponding rise and fall time. The idea is illustrated in figure 2.1 and the delay difference is quantified in Table 2.1. Moreover, the strength of the previous driving stage, or gate drive, is only known after the synthesis step and the modern synthesis tools are not able to take into account the voltage drop during the data latching to properly size the previous gate. Therefore, we forbid inputs arriving on a transmission gate but only on capacitive transistor gates which do not add stack.

**Table 2.1:** Computed delay difference between an ideal slope and the driving strength of the previous stage. Depending on the supply voltage, there can be a substantial different of up to 23% on the computed delay.

| Configuration: | 1V X1 | 1V X8 | 0.35V X1 | 0.35V X8 |
|---|---|---|---|---|
| rise time | 19% | 1.2% | 22.3% | 2% |
| fall time | 18% | 1.4% | 23% | 1.8% |

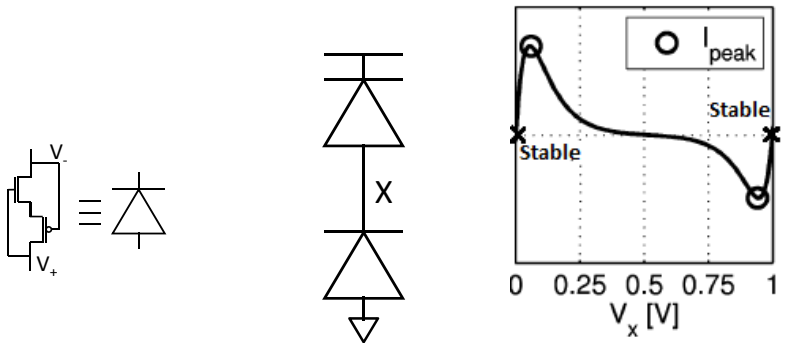
#### *2.2.1.3 Scan, reset and inverted output functionalities*

Testing determines the yield of a given circuit and gives indications on how to improve it. With the complexity of modern ICs, it is impossible to have pins with direct access to all the desirable parts of the circuit. Therefore, designers use one or several scan chains as an artificial mean to access the circuit inside and show the correct behaviour of the circuit, and it is today compulsory in advanced designs [58]. The principle of this technique is the following: scannable flip-flops are used in the functional datapath with two inputs - data in (D) and data test in (TI) - selectable by a test enable signal (TE). During the automatic Placement&Route (P&R), the outputs (Q) and the test inputs (TI) are connected to each other to form a chain of flip-flops called scan chain. If the length of the chain is N, known values of data are fetched in the circuit during N clock cycle. At the end, the state of the circuit is perfectly known, so is the expected output of the combinational logic. Thanks to that, the outputs of the chip are detected and compared to the expected outputs.

As the test represents a lot of time in the industrial process, the trend is to reach similar functional and scan frequencies. Thereby, the scan path and the D-to-Q path will be similarly sized in this work.

Processor designers intensively use the reset function in order to know precisely the initial state of the circuit. Synthesises of industrial state of the art processors, namely TI MSP430 and ARM Cortex M0, were performed and have been shown that the majority of the flip-flops is resettable and with minimum drive. Consequently, the reset function will be also added to our architecture in order to compare realistic flip-flops.

As we could see in Section 1.3, flip-flops may have a different number of stages between the input and output. It means that the direct and shorter output might also be the complementary incoming data. To compare the best performances of the different flip-flops, an inverted output is chosen if it minimizes the D-to-Q delay with the previous constraints and it is considered that the synthesis tool will handle it in the following combinational datapath. The choice is the same for negative triggering edge.




**(a)** The transistor schematic of the ultra-low power diode.

**(b)** The bistable element composed of two ultra-low power diodes.

**(c)** Principle of the current-voltage characteristics of the bistable.

**Figure 2.2:** Ultra-low power static bistable element [59].

Finally, let us sum up the design constraints chosen in this work before developing the latch architectures:

- Static CMOS logic,
- No short circuit path, especially during writing,
- A single transistor pass-gate is forbidden,
- A stack of four transistors is also forbidden,
- The flip-flops are single-edge-triggered,
- The output of the flip-flop must be the output of an inverter,
- The external inputs all arrive on a transistor gate,
- The flip-flops are scannable and resettable,
- An inverted output and negative triggering edge are allowed.

Keeping these constraints in mind, we will develop in the next section the kind of latch that would be interesting for the targeted energy-efficient applications.

### 2.2.2 Architectures of static R-S pulsed-FFs

In this section, the architectures of the static resettable and scannable pulse-triggered flip-flops are elaborated. After motivating the choice of the bistable element structure, different possibilities of writing systems are discussed and the most promising energy-efficient ones will be kept for the comparison.

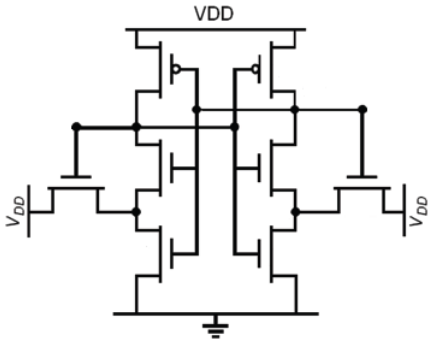



**Figure 2.3:** The half Schmitt trigger bistable [61].

*Bistable element*

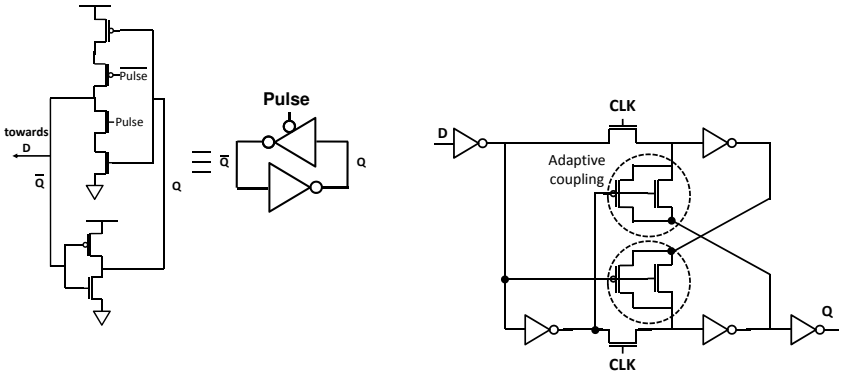
A static latch needs at least one bistable element. Three radically different architectures are proposed in the literature :

- The conventional cross-coupled inverters,
- The ULP latch [59, 60] (figure 2.2) ,
- The Schmitt trigger, either original or simplified [61] (figure 2.3).

The ULP latch consists in two ULP diode in series and interestingly exhibits only one logic node. As the current-voltage characteristics in the combination of two ULP diodes [59], this configuration provides two stable points (see figure 2.2c). The maximum current of an ULP diode is the drain current of a transistor with $V_{GS} = 0$, as shown in figure 2.2a. Consequently at nominal voltage operation, every capacitive coupling will change the voltage value and the recovery time to reach a supply value is several orders of magnitude lower than the two cross-coupled inverters. Therefore, it would impact the driving strength of the inverted output and is thus not suitable for nominal $V_{dd}$ operations [62]. At ultra-low voltage, the leakage current of the writing system has a huge impact on the static voltage value. Indeed, if the writing system is not in ULP logic, the off-current will be in the same order of magnitude than the on-current of the ULP bistable. A careful and complex sizing study must so be performed to ensure functional operations in every environmental cases [60].

The simplified Schmitt trigger was proposed in [61] to increase the yield of the SRAM cells to the detriment of the area. As the read operation is the first cause of failure of 6-transistor SRAM cell, [61] modifies the bistable and increases the read signal to noise margin (RSNM) thanks to a half-Schmitt trigger in the push-down system of the bitcell.

Nevertheless, the topology of static CMOS flip-flop is not subject to any RSNM because the reading is performed without any perturbation in the bistable




**(a)** Breaking feedback mechanism in a latch. **(b)** The adaptive coupling pulsed latch (ACPL) with input data as breaking signal (inspired by [48]).

**Figure 2.4:** Breaking feedback mechanism in a cross-coupled inverters bistable.

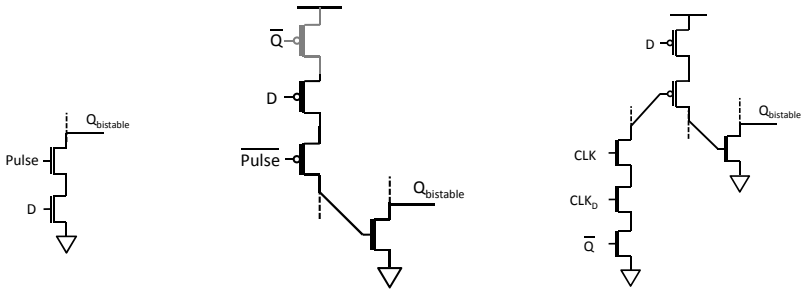
element. This technique is so not relevant in our case. In conclusion, the conventional cross-coupled inverters will be used as bistable element of our latch.

### *Breaking feedback mechanism*

As the conventional bistable element is selected, the conventional technique of breaking bistable feedback (figure 2.4a) can be used to fulfil the constraint of none short circuit path. Let us note the signal breaking the feedback may not only be the clock or pulse signal. Generally speaking, it is any signal which enables the pull-up or pull-down system.

A feedback mechanism using the input data instead of the clock signal was proposed in [48]. The advantage of increasing the data load instead of the clock load is that the data activity is usually much lower than the clock activity (namely 100%), leading thus to a reduced energy consumption. Nevertheless, the evolution of the data input during the whole clock cycle is basically random. Thus, if the data value changes slightly after the triggering clock edge, a node of the bistable element becomes floating and we get every drawback of dynamic logic. That is why only the master latch uses this technique in [48] and why the hypothetical adaptive coupling pulsed latch (see figure 2.4b) will be discarded from the comparison of this work.

Consequently, the writing system will be connected to only one node of the bistable element and, by constraints, at least two transistors will be used to perform both the pull-up and pull-down systems.




**(a)** Writing system with only one stage. **(b)** Writing system with two stages. **(c)** Writing system with three stages.

**Figure 2.5:** Schematic principle of the writing system with many stages. Only the pull-down part is shown.

*Writing system*

Generally speaking, the writing system has to pull down the output value of the flip-flop $Q$ when

$$Pulse = 1 \quad \& \quad D = 0 \quad \& \quad Q_{prev} = 1 \tag{2.1}$$

and has to pull up when

$$Pulse = 1 \quad \& \quad D = 1 \quad \& \quad Q_{prev} = 0 \tag{2.2}$$

This logic equation can be implemented with a stack of three transistors having respectively $Pulse$, $D$, and $Q_{prev}$ on their gate. But, the use of the third transistor is useless because the information of $Q_{prev}$ is already contained in the node of the bistable. Thus, a stack of two transistors with $Pulse$ and $D$ on their gate is sufficient (figure 2.5a). If the $D$ gated transistor is directly connected to the bistable node and not to the supply nodes, a glitch of the input during the clock cycle will charge or discharge an intermediate node of the stack. As it will result in an useless energy consumption, the $Pulse$ gated transistors must be directly connected to the bistable and the $D$ gated to the supply voltage. The drain of the $Pulse$ gated PMOS and NMOS can be connected to form a pass-gate (TGPL configuration) or let separated ($C^2$MOS configuration).

Let us analyse the case where only one PMOS and/or one NMOS achieve(s) the pull-down or pull-up system(s).

In both cases, the transistor has to be off when $Pulse = 0$ and becomes in on-state only if a new value must be written, meaning that the signal on the gate brings the information. For example, a stack of two PMOS, with $Pulse$ and $D$, can be connected to a NMOS, as drawn in black in figure 2.5b. When $Pulse = 1$

& $D = 1$, the PMOS stack charges the intermediate node, the NMOS becomes on and discharges to ground the internal node of the bistable element containing the output information ($Q_{bistable}$). To avoid a discharge of the intermediate node when $Q_{bistable}$ is already at 0, and thus to increase the energy-efficiency, the preceding stack may contain the $Q_{prev}$ information thanks to a third transistor as illustrated in 2.5b. Let us notice that this configuration is the complementary of the CDFF architecture presented in [46]. On the other hand, the pull down system is a stack of two NMOS directly connected to the bistable element. It needs the complement input $\overline{D}$ but avoids stacking three PMOS known as the slowest transistor configuration in this technology (for identical threshold voltages).

Following these ideas, another possibility to avoid a stack of three transistors, is a stack of two transistors commanding the last stage connected to the bistable and containing the entire information $Pulse$, $D$, and $Q_{prev}$, as shown in figure 2.5c. The question is: in which order should they be ? If the gate signals of this stack are $Q_{prev}$ and a signal commanded by $Pulse$ and $D$, it will increase the D-to-Q delay and produce switching energy overhead in the first stage when $D = Q_{prev}$. If the gate signals of this stack are $Pulse$ and a signal commanded by $Q_{prev}$ and $D$, it will increase the D-to-Q delay and produce switching energy overhead in the first stage if $D$ changes during the clock signal. If the gate signals of this stack are $D$ and a signal commanded by $Q_{prev}$ and $Pulse$, the D-to-Q path is not augmented and there is no switching energy overhead in any stage. That is why it is the choice of [49]. As a XOR gate does not bring the information of one or zero but only compared $D$ and $Q_{prev}$ signals, [49] uses pseudo-NAND and pseudo-NOR gates. The pseudo-NOR gate needs a stack of three PMOS but the timing performances are not under concern in the pulse generation.

A four stages configuration is basically the CP$^3$L configuration without the $Q_{prev}$ signal in the pulse generator, which would lead to an useless energy consumption.

After the triggering edge, the bistable element must see the writing system as a high impedance. If there is only one transistor in the pull-up or pull-down system, its gate voltage must come back to its initial value after writing. This *precharge* mechanism can be implemented in different ways. An extended and applied discussion will be performed in Annex A with the CDFF architecture.

### *Scan and reset functionalities*

As a result from the previous sections, the architectures adopted for the comparison are the transmission gate pulsed latch (TGPL), the complementary CMOS (C$^2$MOS), the conditional discharge flip-flop (CDFF) and the conditional push-pull pulsed latch (CP$^3$L).

For pulse-triggered flip-flops, the reset function is easily implemented by adding two transistors in the bistable element, one to pull a node to *gnd* or $V_{dd}$ and one to break the feedback. At least one other transistor must be added in the writing system to ensure no short current path at triggering clock edge.

The scan functionality is implemented with minimum additional transistors, the size of which is the same as the functional data path to ensure a scan test in every operating condition of the circuit (see section 2.3.1).

For TGPL and $C^2$MOS pulsed-FFs, there are two ways to elaborate the scan function (figures 2.6 and 2.7): either the data stack is enlarged then duplicated, having so a basic CMOS multiplexer with the test enable signal (TE) as command signal and D and TI as multiplexed inputs, or both data and clock stack are duplicated, presenting two AND gates in the pulse generator but a lower stack in the latch. Let us notice that the TI-to-Q path of the -Clk architectures can be connected by the other node of the bistable element ($Q_{int}$). It would reduce the junction capacitance at the intermediate node and thus decrease the D-to-Q delay. Nevertheless, in this configuration, the TI-to-Q path presents one additional stage compared to the D-to-Q path. If we want to perform scan test at ultra-low voltage, the delay generated in the pulse generator has to be ensured large enough for this worst case and we will see in the next chapter that guaranteeing a sufficient delay at ULV is a key bottleneck for pulsed-FFs. As we consider that scan test must be available at ULV, this option is discarded in this work.

For CDFF architecture (figure 2.8), the data stack may not be enlarged because a stack of four is forbidden. Thus, the stack must be duplicated as well as the precharge part.

The $CP^3L$ architecture on figure 2.9 is the -Data version obtained in the same way of TGPL and $C^2$MOS architectures. The $CP^3L$-Clk architecture was discarded because presenting a too important amount of transistors.

The pulse generators of each latch have the same sizing strategy, *i.e.* achieving an FO3 slope on the clocked transistors. It allows us to take into account the real cost of an increase in speed for the energy consumption.

Finally, the schematic of each selected static resettable and scannable pulse-triggered flip-flop architectures are shown in figures 2.6 to 2.9. Let us notice that one of the differences between these architectures is the number of stages in the D-to-Q path. Without considering the output inverter, we can differentiate the one-stage (TGPL and $C^2$MOS) and two-stages (CDFF and $CP^3L$) topologies.

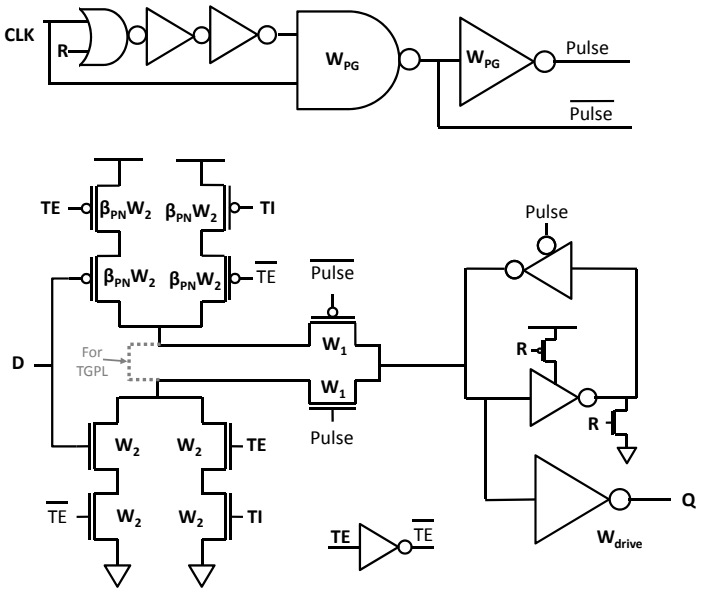



**Figure 2.6:** One-stage TGPL-Data and C$^2$MOS-Data (with and without the doted wire, respectively) FFs, two variants of the TGPL architecture from figure 1.15.

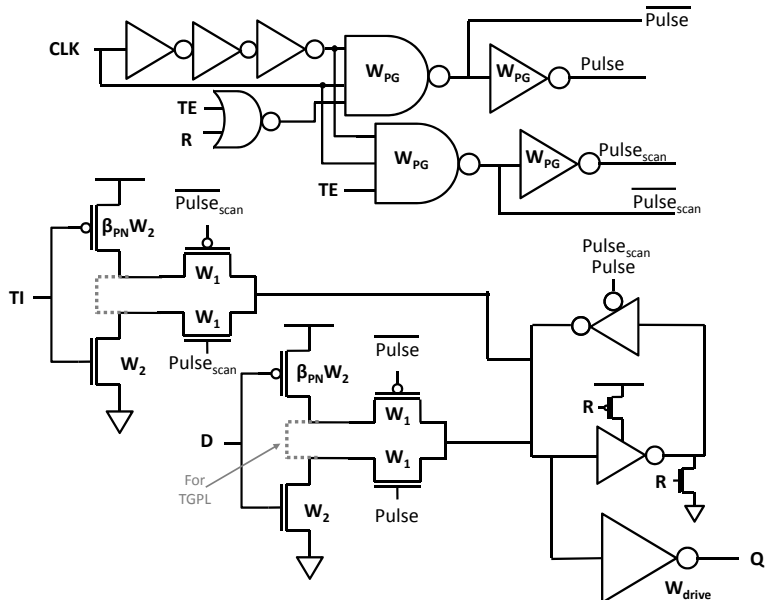



**Figure 2.7:** One-stage TGPL-Clk and C$^2$MOS-Clk (respectively with and without the doted wire) FF.s, two variants of the TGPL architecture from figure 1.15.

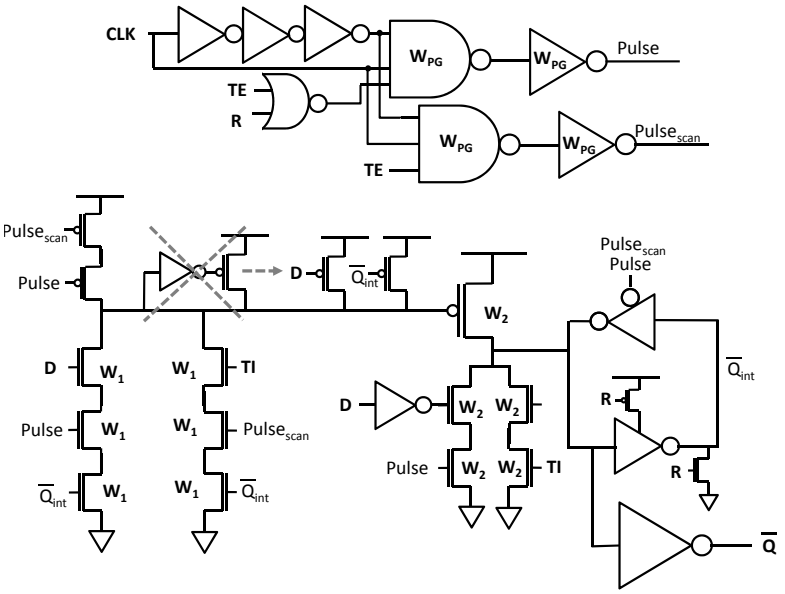



**Figure 2.8:** Scannable and resettable version of the two-stages single-edge CDFF from figure 1.17d

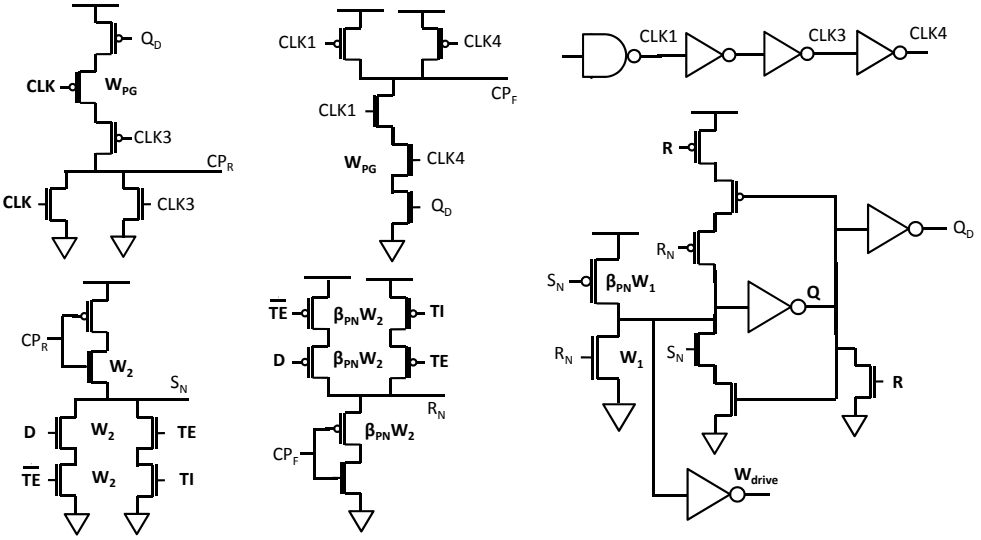



**Figure 2.9:** CP$^3$L-Data, scannable and resettable version of the two-stages negative-edge triggered CP$^3$L from figure 1.21.

## 2.3 COMPARISON AND RESULTS

This section starts by explaining the sizing methodology used in this work, focusing on the $\beta_{PN}$ ratio of transistors and the concept of energy-efficient curves (EECs). Then, the results of the comparison of all FFs are shown at nominal (1V) and optimal energy (0.35V) voltage. The conclusion of the comparison is then discussed and the C$^2$MOS-data and TGPL-Clk topologies are pointed out.

### 2.3.1 Sizing methodology

The proper ratio $\beta_{PN}$ between the gate width of the PMOS ($W_{g,p}$) and the gate width of the NMOS ($W_{g,n}$) in a CMOS stage, depends on the design target. Generally speaking, the optimum $\beta_{PN}$ is not the same if equal rise and fall delays or transition times are on purpose or if the energy or robustness is under consideration [11, 63, 64, 65]. Moreover, the stack is another parameter for the sizing.

For circuits working at nominal voltage, the timing properties are of primary importance and equal rise and fall times are targeted. For circuits working at ultra-low voltage, the ratio between the on- and off-current in a CMOS stage is the main consideration as it will be shown in Section 3.4. But the choice is more difficult in UWVR circuits where the logic works both at nominal voltage and ULV. Spice simulation have been performed at 1V and 0.35V for a stack of one and two transistors and the optimum $\beta_{PN}$ ratio for timing consideration and robustness are shown in Table 2.2.

**Table 2.2:** Optimal $\beta_{PN} = \frac{W_p}{W_n}$ ratio for identical rise and fall time at nominal voltage and for maximal $\frac{I_{on}}{I_{off}}$ ratio at ULV.

| CMOS stage stack | $\beta_{PN}$ at nominal voltage (identical rise/fall times) | $\beta_{PN}$ at ULV (maximal $\frac{I_{on,lin}}{I_{off,sat}}$) |
|---|---|---|
| 1 | 1.6 | 0.38 |
| 2 | 2 | 0.37 |

In FDSOI 28nm technology with LVT feature and nominal back biasing (vdds=gnds=0V), the optimal $\beta_{PN}$ ratio for a timing match is about 1.6 for a stack of 1 and 2 for a stack of 2, both at 1V and 0.35V. The fact that results are identical for 1V and 0.35V, is particular to the 28nm FDSOI technology. The subthreshold current per unit width is actually higher for the PMOS transistor than for the NMOS in this technology. It is explained by the fact that PMOS have a lower threshold voltage to compensate the loss of mobility at nominal voltage. And indeed, despite their higher leakage current, PMOSs need a larger $W_g$ to provide an identical timing at nominal voltage.

Those observations may lead to a lot of discussions about the proper sizing to use in a circuit working both at nominal and near- or subthreshold voltage. In this work, we chose to target identical timing performances at nominal voltage and assume that the robustness problem at ultra-low voltage can be handled in FDSOI thanks to the proper choice of back bias values (see Section 3.4).

*Energy-efficient curves*

FF architectures can be thoroughly and fairly compared by extracting the energy-efficient curve (EEC) in the energy-delay ($E-D$) space [6]. This curve is the set of design points showing minimum energy (delay) for a given delay (energy) [66]. From the theory, EEC has a hyperbolic shape and allows the understanding of the $E-D$ tradeoff of FF in both high-speed and low-energy designs. Our sizing optimization methodology is largely inspired from [6] and the details of the testbench can be found in Annex B. Therefore only the outlines are presented in this section.

Only the gate width of the transistors in the D-to-Q path can modify the speed and thus the energy-delay tradeoff. Thereby, those gate widths - called $W_k$ ($k = 1, 2, ...$) in figures 2.6, 2.7, 2.8, and 2.9 - are the main variables for the transistor sizing algorithm. Once the sizing variable $w_k$ are chosen, it is possible to apply an optimization algorithm to extract the optimum $w_k$ for each point of the EEC. As already mentioned, the EEC is made up of the design points minimizing the $E^iD^j$ figures of merit (FOMs). The exponents $i$ and $j$ are predetermined integers balancing the contributions and a particular $E^iD^j$ FOM is a choice of the designer on the energy and timing characteristics of the circuit. In the neighbourhood of the design minimizing a given $E^iD^j$ FOM, a $j$% performance increase is traded for a $i$% energy increment and vice versa [67]. Thus, the designers have to choose which figure of merit is targeted for their application.

**Table 2.3:** Imaginary example of sizing algorithm results.

| FOM to minimize (Designer choice) | $W_k$† (Algo. result) | corresponding $D$ (Algo. result) | corresponding $E$ (Algo. result) |
|---|---|---|---|
| $E^2D^1$ | $1W_{g,min}$ | 6 | 1.5 |
| $E^1D^1$ | $2W_{g,min}$ | 4 | 2 |
| $E^1D^2$ | $3W_{g,min}$ | 3 | 3 |

†: Logically for realistic designs, the higher the gate width is, the lower the delay and the higher the energy are.

For a given set of $W_k$ which minimizes a particular $E^iD^j$ FOM, it is associated a propagation delay and an energy consumption of the flip-flop. Hence, each design point can be plotted in the energy-delay space ($E-D$) as represented in

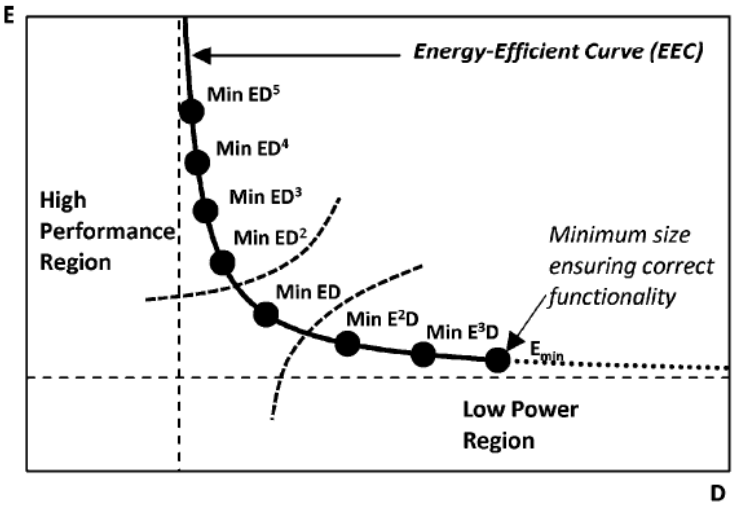



**Figure 2.10:** Extraction of energy-efficient curve (EEC) [67]: from the corresponding energy consumptions ($E$) and propagation delays ($D$) of the design points given by the $W_k$, we plot them in the $E - D$ space and interpolate those points by a hyperbole.

figure 2.10. As the theory also tells us that the EEC has an hyperbolic shape, it is possible to select a discrete set of the $E^iD^j$ FOMs and interpolating them to extract the intermediate points. This allows to get the energy-efficient curve without applying an optimization algorithm for all the EEC points, which would take a infinite amount of time.

In this work, we chose to consider the FOMs $ED^j$ and $E^iD$, for $i, j = 1...5$, because they cover a very wide range of applications. After having determined the nine sets of $W_k$ which provide an optimal design point for each FOM, we got nine couples of energy consumption and delay propagation which correspond to the nine design points. Then, we plot those couples in the energy-delay space, as in figure 2.10, and we interpolate this set of nine points by a hyperbole. The obtained hyperbolic curve is the EEC of a given flip-flop and represents the corresponding $E$ and $D$ of the design points minimizing all the $E^iD^j$ FOMs. In other words, this curve shows us what the cost in energy for a given propagation delay is. Thanks to the EECs, we are able to fairly compare the FF architectures over a wide range of FOMs, thus applications.

In order to provide technology independent results, delay and energy are normalized respectively by :

- $D_0$ the propagation delay of a Fan-out 4 (FO4) inverters chain,
- $E_0$ the energy dissipated during a complete switching cycle of a minimum sized symmetrical inverter without output load.

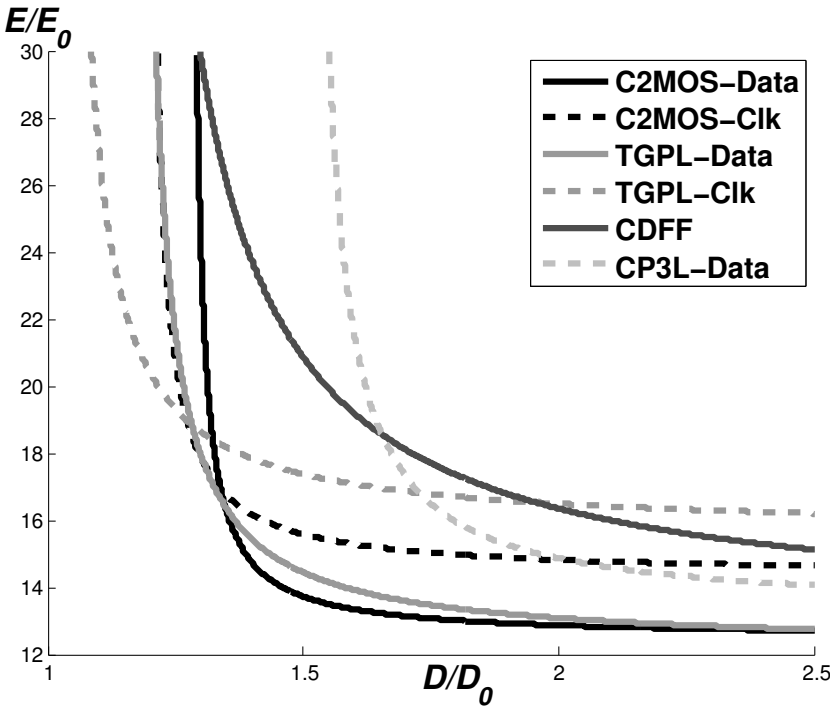



**Figure 2.11:** EEC: $V_{dd}$ = 1.0V, vdds=gnds=0V, $\alpha_{sw}$ = 0.15, $T_{clk}$ = $40FO4$, corner TT, temperature 70°C.

Let us notice that the most significant difference with the testbench of [6] is the use of a novel setup time metric dedicated to pulse-triggered flip-flop [28]. This metric is focused on the output transition of the flip-flop during a sweep of the input, instead of the Clk-to-Q or D-to-Q delay as it is the case in the literature and the industrial characterization tools. We showed that, for pulsed-FFs, it provides the most timing-efficient setup time considering the whole pipelined CPU. Finally, let us remind that the propagation delay of FF is defined at the D-to-Q time when the input data $D$ arrives at the setup time, defined as a variation of 10% of the output transition time (see [28] for more details).

### 2.3.2 Nominal voltage operation

The energy-efficient curve of all the scannable and resettable pulsed-FFs are reported in figure 2.11, for a nominal supply voltage and a minimum driving strength (X1). The temperature is 70°C, the data switching activity factor ($\alpha_{sw}$) is 15% and the clock period ($T_{clk}$) is 40 times the propagation delay of FO4 inverters in the same environmental conditions. All those values are typical for the

targeted applications and do not heavily impact the outcomes of the comparison [6].

We see that the *-Data architectures provide the lowest minimum energy consumption, with C$^2$MOS slightly better than TGPL. During a C$^2$MOS transition, the intermediate nodes of tristate input inverters are not completely charged or discharged, because the gate to source voltage of the first transistor in the stack decreases gradually below the threshold voltage. Therefore, C$^2$MOS architecture saves the dynamic energy of fully charging and discharging the junction capacitances proportional to $W_2$ (see figure 2.6). This lower dynamic consumption allows C$^2$MOS architectures to get higher $W_k$, so higher speed, for the same energy. Nevertheless, C$^2$MOS are finally outperformed by TGPL topologies in high-speed region, since the two transistors of the transmission gate help to improve the transition speed. *-Clk architectures provide a better $E - D$ tradeoff in the high-speed region, due to their smaller stack in the input stage. Let us notice that two-stages structures (CP$^3$L and CDFF) presents a poorer tradeoff on the overall $E - D$ space.

### 2.3.3 Ultra-low voltage operation

The EECs of all the pulsed-FFs, for a supply voltage of 0.35V (minimum energy per operation in 28nm FDSOI technology) and a minimum driving strength, are reported in figure 2.12. Here the operating temperature is 25°C because the self-heating of ultra-low power circuits is almost negligible and those circuits normaly work in ambient temperature.

We see that the C$^2$MOS-Data architecture is the most energy-efficient in almost the whole $E - D$ domain, excepted in very high-speed region. Thanks to the stack, this architecture exhibits a very low leakage current which becomes extremely significant in ULV operation and so dramatically impacts the energy consumption.

At ULV, TGPL architectures are less energy-efficient over the whole $E - D$ space. While they are keeping a higher leakage current as C$^2$MOS, their benefit in timing is jeopardized by the sub-threshold regime. Indeed, during a transition, the $V_{GS}$ of one of the two transistors of the transmission gate progressively decreases. As it exponentially depends on the $V_{GS}$ in sub-threshold domain, the drain current of one of the two transistors of the transmission gate is negligible a short time after the beginning of the transition. It leads to a C$^2$MOS-like topology, with higher diffusion capacitances in the D-to-Q path.

Again, the two-stages architectures are less energy-efficient than the one-stage ones in every region of the $E - D$ domain. The fact that CDFF and especially CP$^3$L structures present the worst energy-delay products (EDP), in opposition to the results in [49], shows that the reset and particularly the scan functionalities may change the conclusion of comparisons. Therefore, it proves that the facility of implementing the scan function must be taken into account in the choice of FF architectures.

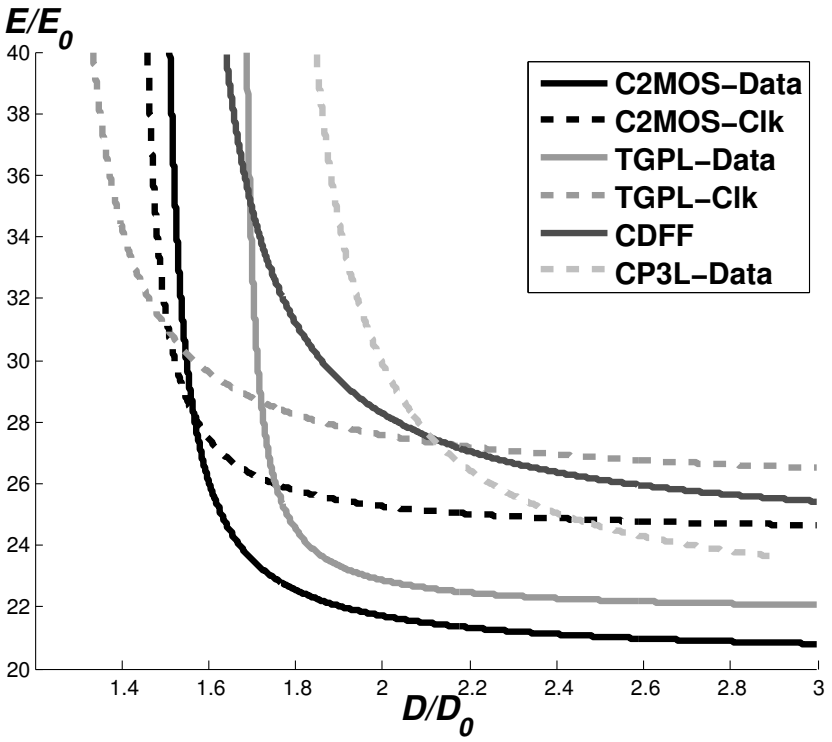



**Figure 2.12:** EEC: $V_{dd}$ = 0.35V, vdds=gnds=0V, $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, corner TT, temperature 25°C.

### 2.3.4 Back-biasing technique

In the previous section, we saw that C$^2$MOS-Data architecture is the most energy-efficient topology for energy-efficient and low-power applications at ultra-low voltage and nominal voltage. On the other hand, TGPL-Clk exhibits the lowest $ED^n$ products for $n \geqslant 2$ and is thus dedicated to high-speed applications.

These conclusions apply to the 28nm FDSOI technology but, generally speaking, there might be more than two optimal architectures for a large set of $E^iD^j$ FOM. If an application needs both a high performance on critical path and a low power consumption for the circuit, the selected architecture needs to be designed with all possible sizing to reach a pareto optimum-like energy-delay tradeoff on the whole circuit. Nevertheless, it means that the architectures must be properly sized for each figure of merit targeted and all characterized in each PVT conditions for all the drives provided in the library. All this may significantly increase the design and computation time.

This tradeoff between design time and energy-efficiency can be almost fully alleviated thanks to the FDSOI technology. Figures 2.13 and 2.14 compare the

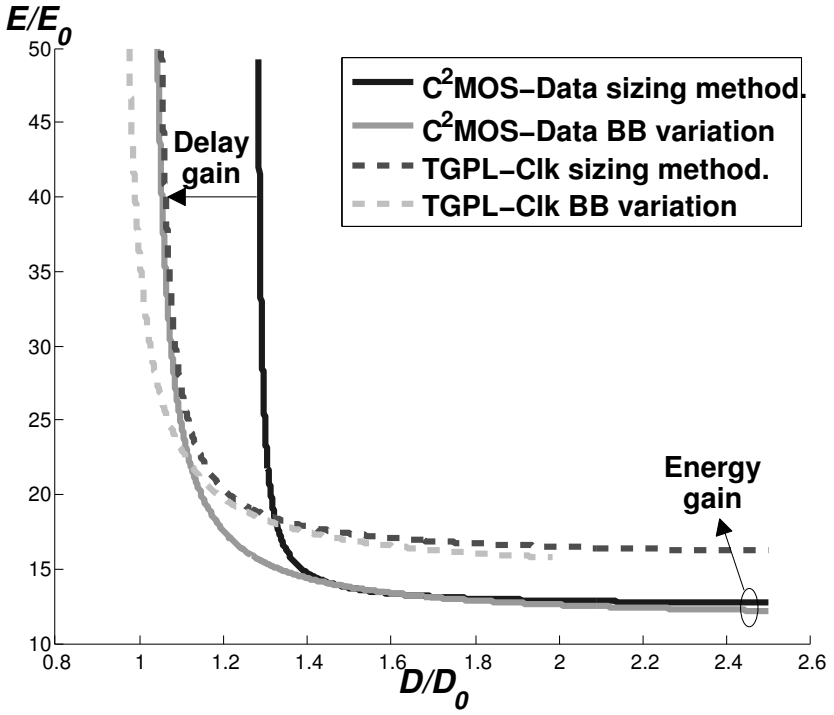



**Figure 2.13:** EECs of the C$^2$MOS-Data and TGPL-Clk architectures extracted from the sizing methodology and from a applied back biasing ($V_{dd}$ = 1V, vdds/gnds range = $\pm$ 1V , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, temperature 70°C). Back biasing technique provides better performances than sizing methodology.

performances in the $E - D$ space of the C$^2$MOS-Data architecture (minimum $ED$ sizing) on which a wide symmetrical back biasing is applied. Figure 2.13 also shows the TGPL-Clk architecture with minimum $ED^2$ sizing, as it is the most energy-efficient in high-speed region. As we can see, the delay in the high-speed region is lower for the same energy than the delay obtained by the sizing methodology. Similarly in the low-power region, the energy consumption is lower for the same delay than the energy consumed by the minimum sizing of transistor. At ultra-low supply voltage (figure 2.14), the impact of the threshold voltage on the on-current is almost as strong as on the off-current. Therefore, an increase (decrease) in delay implies a decrease (increase) for the leakage current. The leakage energy is this leakage current integrated over the clock period that is related to the gates delay. As the transistor dimensions and the supply voltage remain the same, the dynamic energy variation is only due to short-circuit current. Consequently, the energy per operation does not vary that much in the low-power region, *i.e.* for reverse body bias, and remains quite the same as the energy of the

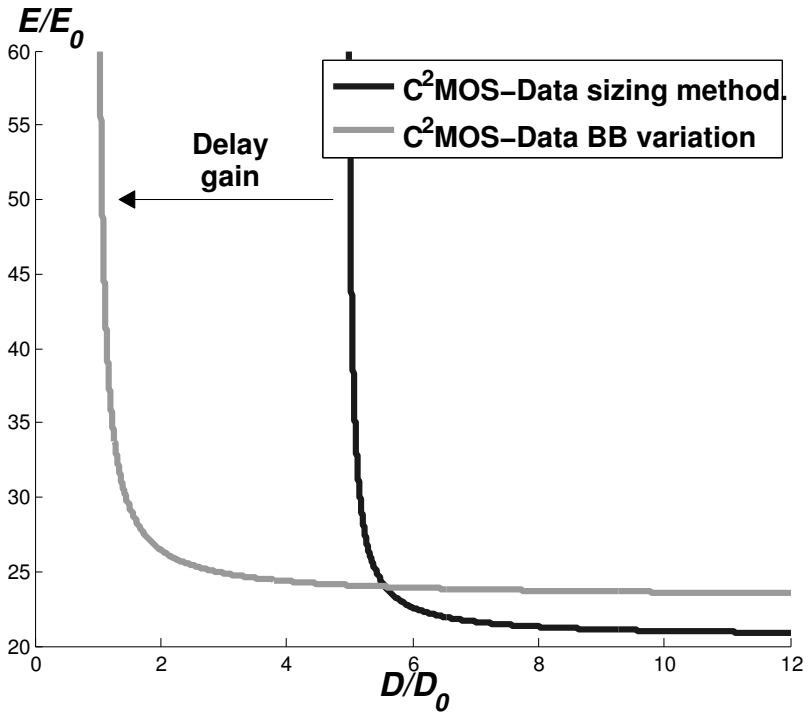



**Figure 2.14:** EECs of the C$^2$MOS-Data architecture extracted from the sizing methodology and from a applied back biasing ($V_{dd}$ = 0.35V, driving strength X1 , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, temperature 25°C).

minimum $ED$ sizing architecture. On the other hand, the relative delay gain is higher at ULV than at nominal voltage because the delay, essentially depending on the on-current, varies much more with the threshold voltage shift. All that clearly shows that we can modulate the energy-delay performances of flip-flops more efficiently than the sizing methodology and also more flexibly because it is not built in hardware and can be dynamically modified during the circuit operation.

## 2.4 IMPLEMENTATION IN DIGITAL FLOW AND SILICON MEASUREMENTS

As we saw in figures 2.6 and 2.7, the C$^2$MOS and TGPL architectures are very similar. The result of their comparison is thus extremely technology dependant. As the design platform of the technology changed during the time of this work,

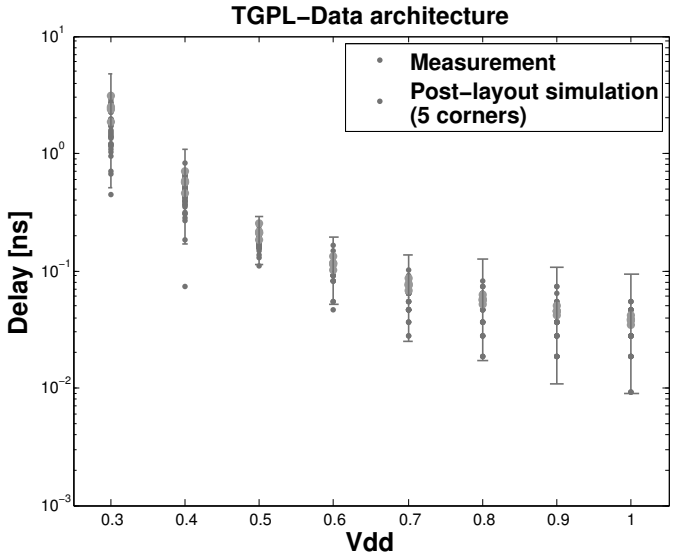



**Figure 2.15:** Measured delays of TGPL-Data from 0.3V to 1V with the expected $3\sigma$ range computed from 63 chips (VDDS=GNDS=0V). Gray dots are the results of post-layout simulations in the five process corners (T°= 80°C).

our first results were that the TGPL-Data architecture was more energy-efficient than the C$^2$MOS-Data, *i.e.* with a lower $E^1D^1$ product.

That is why we selected the TGPL-Data architecture to be implemented on silicon in the high-speed DSP FRISBEE [24]. The TGPL-Data post-layout netlist was characterized by the same automatic tools used for master-slave topologies. To obtain a proper hold time value, we have to set some simulation variables to ensure that the tool does not allow a glitch during the pulse window. Indeed, due to the soft-edge property of the pulsed-FF, it is possible to trigger a right final output value after having seen a glitch at the output. But if this glitch propagates in the paths after the flip-flop, it will lead to an additional energy consumption in the combinational logic.

The setup time of soft-edge FF is obtained by the same way as hard-edge triggered flip-flop, excepted that we showed in [28] that the setup time metric is not optimal. For FRISBEE, the criteria to determine the setup time was an increase of the Clk-to-Q delay of 10% compared to the nominal case. In [28], we showed that this metric, as well as the minimum D-to-Q delay, does not give the limit beyond which the performance of the flip-flop is degraded and the reliability is endangered [68]. A setup time metric based on the output transition time would provide all the timing performances of the TGPL-Data architecture.

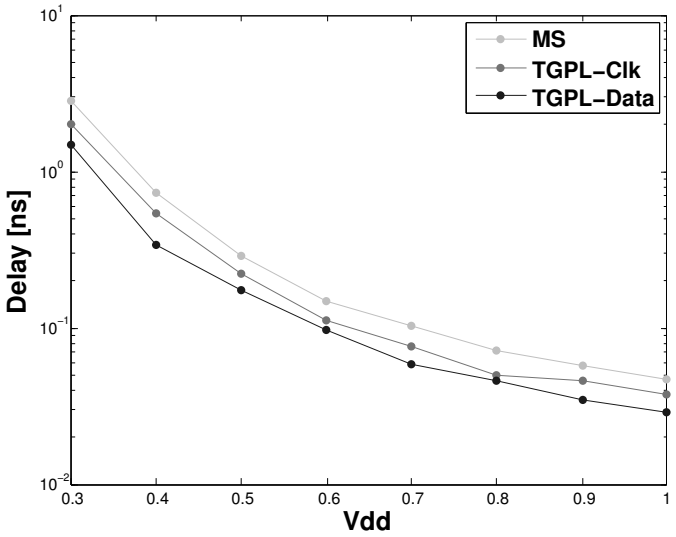



**Figure 2.16:** Measured delays from 0.3V to 1V in nominal case (VDDS=GNDS=0V). Clk-to-Q delay is lower for TGPL-Data than for TGPL-Clk, contrary to the D-to-Q delay (T°= 80°C).

After having obtained the timing characteristics of our pulsed-FF, synthesis and Place&Route were performed without any serious problem for the timing closure. After the signoff verification, the chip was sent in fabrication and the measured performances have expectantly outperformed the state of the art [24].

In parallel, a test circuit was laid out and fabricated to study more precisely the performances of the pulsed-FFs alone. Hence, the following of this section presents post-layout simulations and silicon measurements of TGPL-Data and TGPL-Clk architectures and compares them to the conventional master-slave flip-flop.

### 2.4.1 Silicon measurements

This section presents delays measured on silicon, exhibited by different flip-flops architectures: TGPL-Data with minimum EDP, TGPL-Clk with minimum $ED^2P$, and a conventional master-slave (MS) coming from an industrial standard-cell library. The mean and standard deviation of the delay are extracted from the measurements of 63 chips uniformly distributed on the wafer.

Figure 2.15 represents the $\pm 3\sigma$ dispersion, assuming a Gaussian distribution above 0.5V and a lognormal distribution below, for one architecture at each supply voltage $V_{dd}$. This figure shows that the model used for the simulation is

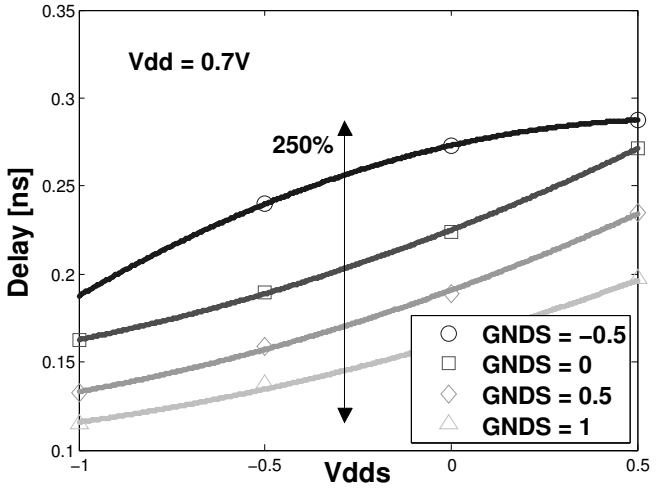
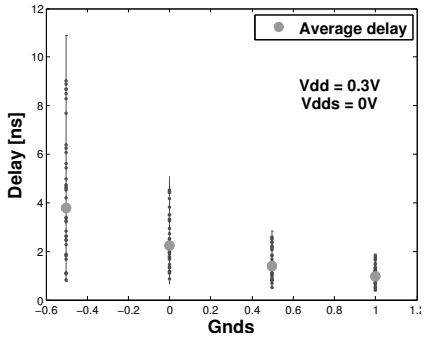



**Figure 2.17:** Vdds variation and GNDS at intermediate supply voltage. A variation of 250% is achieved with the back biasing.

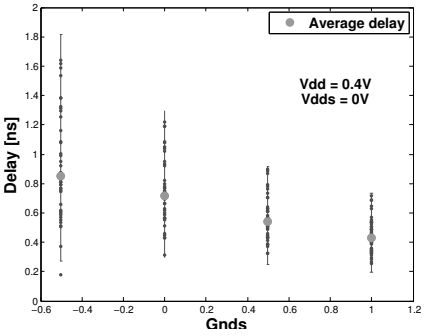
very accurate because the delay computed in the five process corners at each $V_{dd}$ effectively lays in the range of delay coming from measurement. Similarly, it also proved that our measurements are consistent.

We clearly see in figure 2.15 the exponential behaviour of the delay when we go towards the subthreshold regime, *i.e.* an linear evolution in logarithm scale. On the other side, for the highest $V_{dd}$ values, we see a larger delay dispersion ($\sigma$) relatively to the average ($\mu$). This is explained by the accuracy of our measuring equipment. Indeed, as the speed increases with $V_{dd}$, the FFs delay reaches the precision limit of our tester. In other words, the delay dispersion due to the test accuracy becomes identical to the real FF delay dispersion. Consequently, the measured $\sigma/\mu$ ratio increases at nominal voltage.

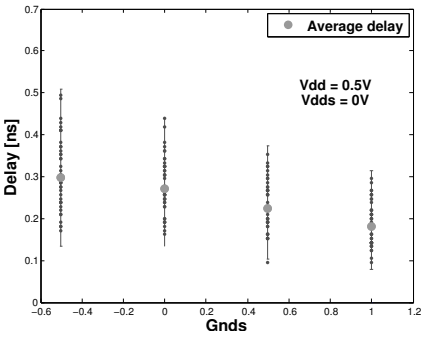
Figure 2.16 compares the measured Clk-to-Q delays of the three selected architectures for a wide range of supply voltage. To remind, the dots represent the average delay computed over 63 chips. As predicted in simulations, the Clk-to-Q delay of the TGPL-Data is faster than the Clk-to-Q delay of the TGPL-Clk. Nevertheless, simulations also showed that the D-to-Q delay, which is the principal timing parameter, is lower for TGPL-Clk than for TGPL-Data. Measurement results showed that this trend is effectively observed but, unfortunately, the measures of the minimum D-to-Q delay are not consistent because of the inaccuracy of the setup time measurement. As quantitative results, TGPL-Clk shows an average Clk-to-Q delay of 31ps at 1V and 80°C, without FBB.



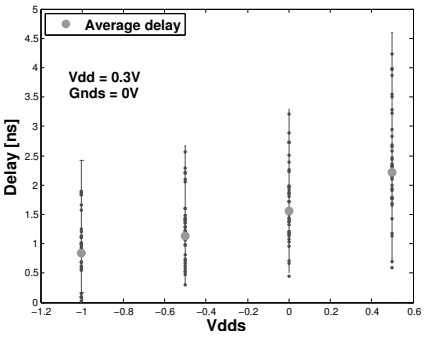

**(a)** Gnds variation at subthreshold supply voltage.




**(b)** Gnds variation at near-threshold supply voltage.




**(c)** Gnds variation at super-threshold supply voltage.

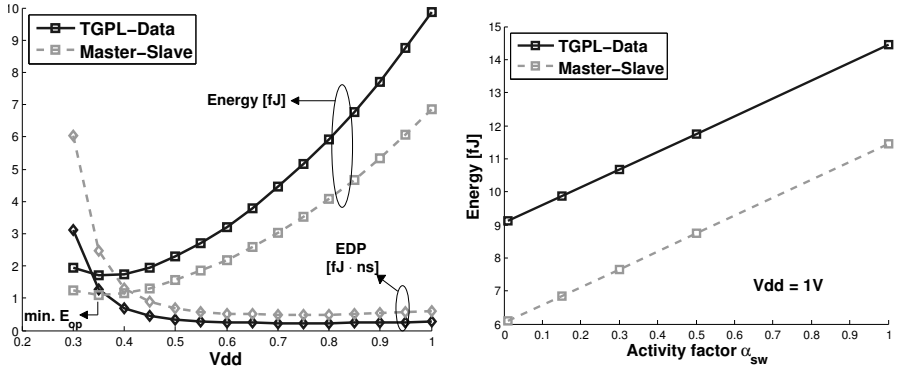



**(d)** Vdds variation at subthreshold supply voltage.

**Figure 2.18:** Evolution of measured delays with VDDS and GNDS.

The above results are for flip-flops in nominal back biasing conditions, namely VDDS = GNDS = 0V. Figure 2.17 represents the evolution of the delay with the back bias for the TGPL-Clk architecture at super-threshold supply. As the flip-flops use LVT transistors, only a strong forward back bias is available. Playing with the back voltages allows to cover a wide range of performance since the maximum delay is 2.5 times higher than the minimum delay. This trend is also visible for the two other architectures. As a reminder, changing the back bias does not affect the dynamic energy.

The back biasing technique has an even deeper impact on the performances when the supply voltage is near the threshold voltage. We can see in figure 2.18 the delay distribution either typical of subthreshold operation (0.3V) or of super-threshold operation (0.5V), depending on the back bias values. In figure 2.18b, the supply voltage is 0.4V, a bit above the threshold voltage. We can




**(a)** Evolution of the average energy per cycle ($\alpha_{sw} = 15\%$) and the EDP with the supply voltage. **(b)** Evolution of the average energy per cycle with the data switching activity factor.

**Figure 2.19:** The TGPL architectures outperform MS topology for all supply voltage between 0.3V and 1V and all activity factors (post-layout simulation).

see that, for a reverse body bias (gnds = -0.5V), the spread of the distribution is subthreshold-like, while for a forward body bias (gnds = 1V), the range of variation becomes closer than the variations at 0.5V. Finally, let us notice that gnds, the NMOS back bias, impacts more on the variation of the delay than the vdds, the PMOS back bias (figures 2.18a and 2.18d).

Finally, Table 2.4 compares the energy performances and efficiency of the three architectures by post-layout simulations. As we can see, the TGPL's architectures, composed of one latch and one pulse generator, consume a bit more than the master-slave architecture. It might surprise since pulse-triggered architectures have only one latch instead of two for master-slave. However, as it will be explained more precisely in Chapter 4, the pulse generator is actually the largest part of energy consumption. Nevertheless, TGPL's architectures are more energy-efficient as soon as we take the delay into consideration. More precisely, pulsed-FFs outperform MS topology for each $E^i D^j$ figures of merit with $j \geqslant 1$ and $i \leqslant 3$, for all supply voltages between 0.3V and 1V and all activity factors (figure 2.19). Let us notice from figure 2.19, that the minimum energy point is effectively at 0.35V and the minimum EDP around 0.7V.

## 2.5 CONCLUSION

This chapter presented our approach to choose the most energy-efficient pulse-triggered flip-flop architecture for our UWVR and ULP applications. First, the corresponding design constraints were exposed and motivated. Then, a theoretical analysis was performed to select the most suitable architectures based on

**Table 2.4:** Comparison of the EDP and ED$^2$P figures of merit for the three architectures (post-layout simulation at 1V and 80°C).

| Architecture | $E_{op}[fJ]$ ($\alpha_{sw} = 15\%$) | EDP [fJ $\cdot$ps] | ED$^2$P [fJ $\cdot$ps$^2$] | Area [$\mu m^2$] |
|---|---|---|---|---|
| MS | 6.72 (ref.) | 1136 (ref.) | 1877 | 4.4 (ref.) |
| TGPL-Data | 10.08 (+50%) | 288 (-74%) | 82 | 5.4 (+23%) |
| TGPL-Clk | 14.88 (+121%) | 334 (-70%) | 74 | 6.7 (+52%) |

the previous constraints. Afterwards, a comparison of the performances in the energy-delay ($E - D$) domain was performed at the nominal and the energy-optimum supply voltages.

It turned out that the C$^2$MOS architecture is the most energy-efficient topology in the low power region at ultra-low voltage and at nominal voltage while the TGPL-Clk exhibits the lowest $ED^{n\geqslant 2}$ products in the high-speed region in both voltages. These conclusions showed that adding reset and scan functionalities may effectively change the result of the comparison. In particular, the scannable and resettable CP$^3$L architecture exhibits a higher energy-delay product than the corresponding TGPL and C$^2$MOS topologies. As different architectures might provide the optimum sizing point for different figures of merit, we highlighted that the wide back biasing range offered by the FDSOI technology allows to almost totally overcome this tradeoff. It has been shown that the flip-flop energy-delay characteristics get better performances with the back biasing technique than with the sizing methodology. In other words, a forward back bias provides a lower delay for the same energy than the architecture sized to reach the very high speed ($ED^5$) figures of merit ; a reverse back bias provides a lower energy dissipation for the same delay than the architecture sized to reach the low power ($E^5D$) figures of merit. Furthermore, the back biasing technique allows to change dynamically the energy-delay performances of flip-flop, on the contrary of a sizing methodology which provides fixed hardware configurations. We are thus allowed to contend that the FDSOI technology is *khtêma es aeí*.

The last section of this chapter presents silicon measurements performed on the most efficient TGPL architectures. First, it is mentioned that the implementation in the chip FRISBEE [24] demonstrated that the pulsed-FFs and their soft-edge property do not raise any additional problem at characterization and timing closure steps. Next, measures confirmed the expected low delay of pulsed-FFs compared to master-slave for a wide range of supply voltage ($0.3V \longrightarrow 1V$) and a relatively high number of chips (63). Results of back biasing technique also showed the great modulation of performances that is achievable with the same hardware sized architecture. The number of tested chips allowed us to exhibit the physical distribution of the delay, showing the sub- or super-threshold regime.

Post-layout simulations showed the energy-efficiency of the TGPL architectures, the EDP of which is unconditionally lower than that of a master-slave topology.

Thanks to these performances, our TGPL architectures are already dedicated to cover a wide range of application, especially where the energy-efficiency and the speed are both of primary importance, like in UWVR circuits. Nevertheless, as shown in Table 2.4, the average energy consumption per cycle and the area are lower for the conventional master-slave flip-flop. Consequently, the master-slave architecture is still commonly used in very ultra-low power applications where the power dissipation is the main target to reduce and the robustness is of primary importance. The next chapter explains how to guarantee the robustness of pulse-triggered flip-flops at ultra-low voltage thanks to the use of the current-starved technique in the delay generator. Later in Chapter 4, the study about the pulse generator will show how the pulsed-FF architectures can finally provide a lower energy consumption and area than master-slave flip-flops.

## CHAPTER 3

# ROBUST AND ENERGY-EFFICIENT PULSE GENERATORS

**Abstract**

After the study of scannable and resettable latches in the energy-delay domain, this chapter analyses the pulse generator with other figures of merit. Firstly, the problems of the pulsed-FFs occurring at ultra-low voltage are exposed and illustrated. It is shown that, so far, designers have to make a tradeoff between the robustness of the pulsed-FF and its energy budget.

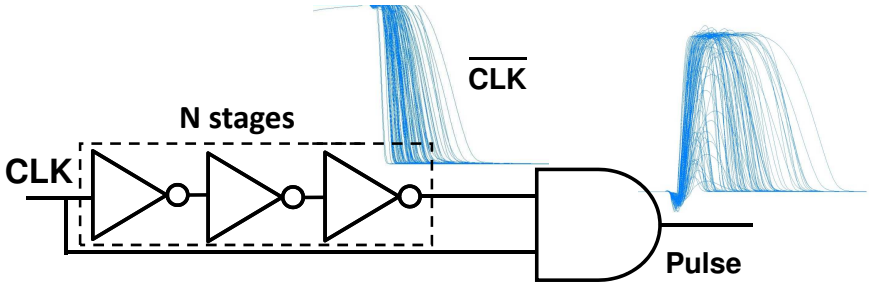
Then, the second section presents and analyses the current-starved (CS) delay generator (DG) which allows to largely avoid the robustness/energy tradeoff. It is shown that the pulse width guaranteeing the correct functionality of the FF, can be modulated without significant increase in the energy dissipation.

Afterwards, a comparison with other delay generators of the literature shows that the current-starved and the conventional DG are the most energy-efficient. Then, post-layout simulations in extreme environmental conditions prove the higher robustness achieved by our delay generator compared to the conventional DG. Therefore, it demonstrates that the current-starved delay generator is the best architecture, regarding the three figures of merit of the delay generators.

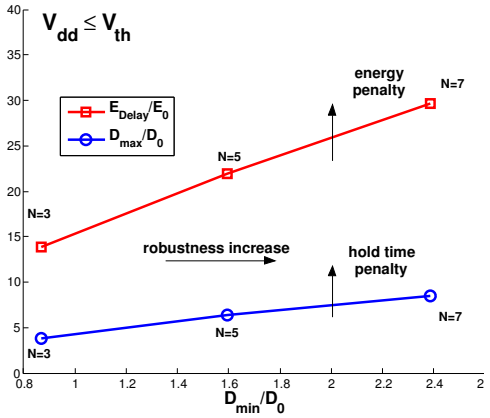
The fourth section analyses the results of silicon measurements. First, we analyse the effects of the back biasing technique on the robustness and show an optimum couple (vdds,gnds) for the yield. Then, the computation of the average yield exhibits the robustness improvement due to the current-starved architecture. Thanks to our proposed DG, an identical yield is achieved at a supply voltage 45mV lower - at ULV - than without the CS technique.

The last section shows how to efficiently add reset and enable functionalities in the pulse generator. As a result, these additional functionalities lead to an increase of 9.7% in the energy-delay product (EDP), while the EDP of the master-slave topology increases by 64%.

**Contents**




**(a)** Variability issues in ultra-low-voltage pulse generator.




**(b)** The energy dissipation and the maximum delay ($\mu + 3\sigma$) versus the minimum delay ($\mu - 3\sigma$) of an inverters chain (N = 3, 5 and 7)

**Figure 3.1:** Increasing the robustness of the pulsed-FF leads to a great cost in energy.

## 3.1 INTRODUCTION

As a reminder, pulse-triggered flip-flops (pulsed-FFs) are made of one latch open during a short period following the triggering clock edge. This period is physically determined by a pulse signal, activating the latch and generated by a pulse generator (PG). The width of this pulse signal is fixed by the delay between the clock and the output of the delay generator (DG) included in the pulse generator (figure 3.1a).

At ultra-low voltage where the impact of local variations is predominant, both the generated delay and the data-to-output (D-to-Q) delay vary significantly from one pulsed-FF to another on the overall circuit. But, with a slow D-to-Q path in the pulsed-FF, the pulse signal can be too narrow to permit the pulsed-

FF to latch the data. A basic solution given to designers is to add stages in the delay chain of the pulse generator (figure 3.1a). Nevertheless, it dramatically increases the hold time and thus decreases the internal race immunity (IRI) [69] of the pulsed-FF. The IRI can be evaluated regarding the maximum pulse width achieved by the PG submitted to local variation. If it leads to a hold time much larger than the mean D-to-Q delay, many delay buffers will be inserted in the short paths. Therefore, an energy overhead is paid twice for inserting additional inverters in the PG, and also delay buffers in short paths in order to fix the hold time violations.

In summary, the pulse window must be large enough to guarantee the correct functionality of the pulsed-FF, and as small as possible to avoid hold time penalties. Consequently, the figures of merit (FOMs) for the pulse generation are

- Minimum delay regarding local variations ($\mu - 3\sigma$ in this work)
- Maximum delay ($\mu + 3\sigma$) or delay dispersion
- Energy consumption

As shown in figure 3.1b, there is a clear tradeoff between the robustness of pulsed-FFs and the energy consumption. To overcome this drawback, we propose a new pulse generator architecture for ultra-low-voltage applications [70]. It is shown that, for an area penalty of three fingers, the robustness is greatly improved compared to the conventional DG used in literature, while the delay spread remains relatively close and the energy consumption is even lower [70].

This chapter is structured as follows. Section 3.2 presents the proposed pulse generator architecture and studies the sizing compared to the three figures of merit. Section 3.3 compares the energetic performances and robustness of a selected sizing of our DG with other architectures. From this comparison, it is shown that the conventional DG and the proposed one are the best candidates for energy-efficient circuits. Then, Section 3.3.1 compares these two pulse generators in post-layout simulations, and exhibits the gain of robustness due to our DG. Afterwards, Section 3.4 studies the minimum operating supply voltage of our FF thanks to silicon measurements. Finally, Section 3.5 shows how adding reset and enable functionalities can be efficiently performed by modifying the pulse generator.

### FOM definition

To evaluate the three figures of merit of the delay generators, we use the same testbench as mentionned in the previous chapter, where the slope is adapted to PVT conditions (the temperature range for our ultra-low-voltage design is from -40°to 85°) and the output load is the clocked transistors of the TGPL-Data architecture with minimum EDP.

The generated delay, the key parameter for the width of the pulse, is defined by the delay between the rising edge of the clock and the falling edge of the

delayed clock signal ($\overline{CLK}$), at 50% of the supply voltage $V_{dd}$. The energy is measured by integrating the supply current over the overall clock period, which is set to the typical logic depth of energy-efficient circuit (40 FO4 delays [67]). For both energy and delay, the mean ($\mu$) and the standard deviation ($\sigma$) of the distribution are extracted from 10000 Monte-Carlo runs. As the delay follows a lognormal distribution in the subthreshold regime [71], the statistical parameters are calculated as

$$\mu = ln(Mean(T_{del})) - \frac{1}{2} ln(1 + \frac{StdDev(T_{del})^2}{Mean(T_{del})^2})$$

$$\sigma = \sqrt{ln(1 + \frac{StdDev(T_{del})^2}{Mean(T_{del})^2})}$$

$$\text{and finally} \quad T_{del,3\sigma} = e^{\mu + 3\sigma}$$

Simulations showed that the energy (dynamic and static together) gets a distribution significantly closer to the normal form than to the lognormal. Thereby, $\mu$ and $\sigma$ are extracted directly from simulations.

As in chapter 2, in order to provide technology independent results, delay and energy are normalized respectively by [67] :

- $D_0$, the delay of a FO4 inverters chain,
- $E_0$, the energy dissipation during a complete switching cycling of a minimum sized symetrical inverter without output load.

## 3.2 CURRENT-STARVED DELAY GENERATOR

Our proposed delay generator architecture is represented in figure 3.2.

The delay generator is a chain of three inverters composed of minimum sized transistors with their source contact connected to two always-on transistors - one PMOS for the pull-up system and one NMOS for the pull-down system. This DG architecture is thus similar to the current-starved ring oscillator architecture in [72]. The size of these current-starving (CS) transistors are called $W_{CS}$ and $L_{CS}$.

The idea contained in this architecture is illustrated in figure 3.3. As we have seen, additional stages are placed in the delay path to increase the minimum generated delay but lead to a wider delay window because of the spread of the generated delay. To alleviate this drawback, our current-starved delay generator architecture exhibits only three stages in the delay path, *i.e.* where CMOS stages are switching during the pulse window, but increases the average generated delay thanks to the CS transistors. As these always-on transistors are not in the delay path, they do not impact (in first approximation) the spread of the generated delay. Hence, this CS architecture can be seen as a translation without expansion of the delay window in the time line. This translation without expansion is the key point to guarantee a sufficient robustness without energy overhead.

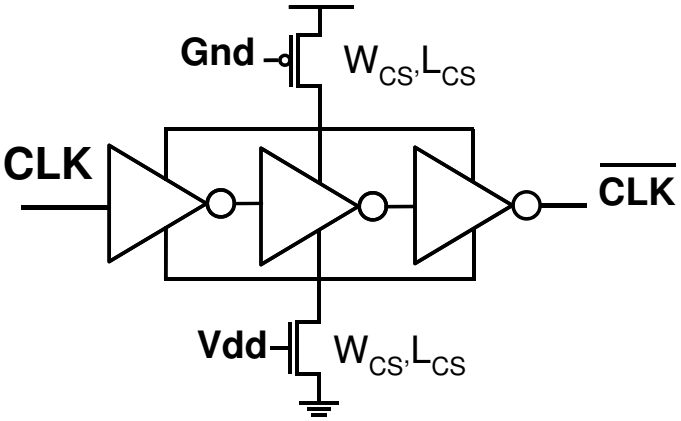



**Figure 3.2:** Proposed current starved (CS) delay generator architecture [70].

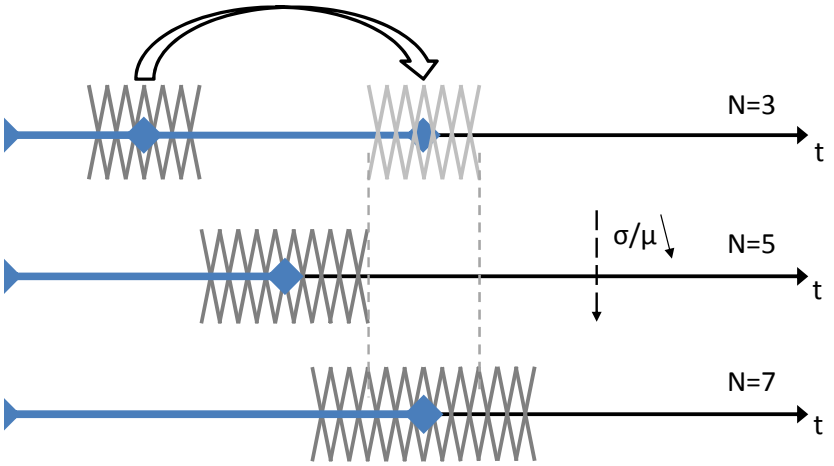
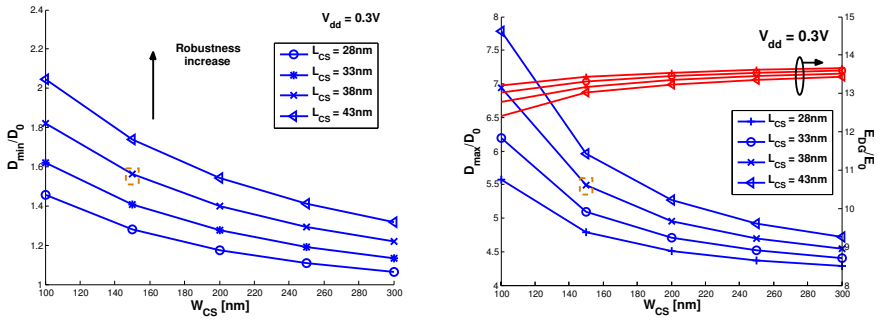



**Figure 3.3:** The current-starved delay generator architecture provides a translation without expansion of the delay window in the time line. In other words, it increases the average delay without noticeably increasing the dispersion.

Figures 3.4a and 3.4b show the evolution of minimum and maximum delays, as well as energy consumption, with $W_{CS}$ and $L_{CS}$ identical for NMOS and PMOS. As expected, a larger gate length provides a lower drain current and thus higher minimum delay. The gate width can be simultaneously increased, in order to reach the same minimum delay with a lower delay dispersion (figure 3.4b). As the gate capacitance of the always-on transistors has no repercussion




**(a)** Minimum delay ($T_{del,-3\sigma}$) function of the size of the current-starving transistors.

**(b)** Maximum delay ($T_{del,3\sigma}$) and energy function of the size of the CS transistors.

**Figure 3.4:** Impact of sizing on the delay generators figures of merit.

on the dynamic energy, the size of current-starving transistors only impacts the leakage energy. As the inverters chain switches at each clock cycle, the switching activity is predominant, and thus the overall energy does not vary as much as delays with gate dimensions. (figure 3.4b).

Let us insist on the fact that the sizes of the current-starving transistors are really powerful and straightforward degrees of freedom. Designers can modulate easily the dimensions of these transistors and reach the desired delay dispersion without impacting the energy consumption.

From this architecture, it directly follows two variants, with only one current-starving transistor (NMOS or PMOS). Figure 3.5 shows that only-NMOS architecture provides a lower delay spread than the only-PMOS one, while both transistors version always provides the lowest dispersion at the same sizes, at the expense of an additional transistor. The energy is not shown because it stays quite the same for the three architectures.

Finally, let us notice that this technique can easily be transposed to dual-edge pulse generators which use a chain of inverters ([73, 74]).

## 3.3 DELAY GENERATORS COMPARISON

Several techniques and architectures of delay generators (figure 3.6) are proposed to designers to increase the minimum delay :

- 3 (the conventional DG), 5 or 7 minimum sized inverters connected in series,
- 1, 2 or 3 inverters with a gate length increase, also called poly bias (PB),
- 1, 2 or 3 half Schmitt triggers (hST) in the delay path for positive triggering clock edge,

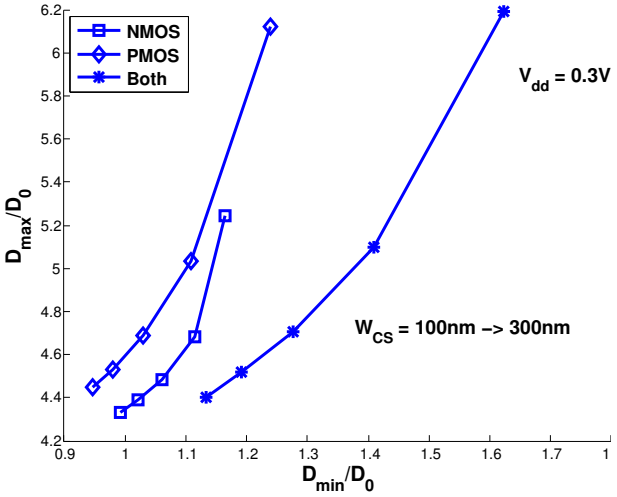
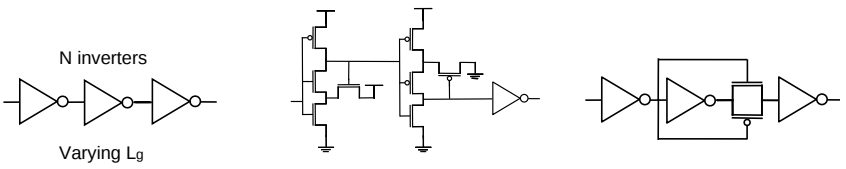



**Figure 3.5:** Minimum and maximum delays ($T_{del,\pm 3\sigma}$) and energy consumption function of the size of the current-starving transistor(s). NMOS-only exhibits a lower timing variability than PMOS-only.

- the insertion of a pass-gate after the second inverter, as proposed in [51].

Delay generator proposed in [39] was discarded because the driving current of the transistors, which are connected to supplies, is, at low voltage, several orders of magnitude lower than the nominal current, leading thus to a generated delay extremely higher than the typical D-to-Q time at the same supply voltage. We have chosen a particular sizing of the current-starved DG, namely $W_{CS} = 150nm$ and $L_{CS} = 38nm$ (surrounded in figure 3.4a and 3.4b), to perform the comparison.




**(a)** Inverters chain: varying stage number and gate length for each stage.

**(b)** Two half Schmitt triggers (hST) placed in the pulling-systems of the delay path.

**(c)** replica delay generator [51].

**Figure 3.6:** Delay generator architectures used for the comparison.

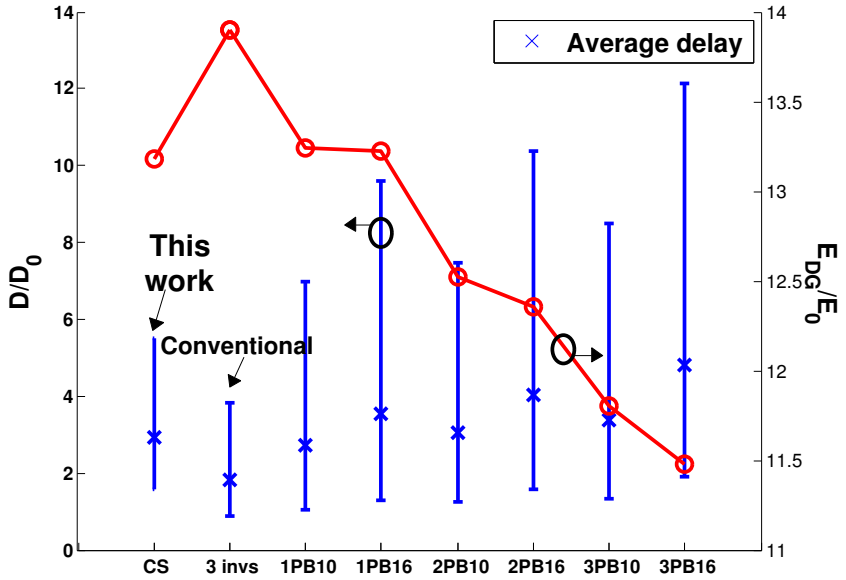



**Figure 3.7:** Comparison of energy and delay performances in subthreshold regime ($V_{dd} = 0.3V$). The name indicates the poly bias and the number of stages getting this PB.

Figures 3.7 and 3.8 show the performances of the delay generators for the three figures of merit. Figure 3.7 compares the current-starved DG with the conventional DG and its variants: one, two or three stages have a PB of 10nm or 16nm. We see that increasing the gate length provides the lowest energy consumption, but it is achieved to the detriment of a huge spread dispersion. Indeed, as the switching energy is predominant in a pulse generator activated at each clock cycle, the saving in energy is quite small compared to the delay dispersion due to the PB. For example, the energy consumption of 3PB16 (the three stages of the conventional PG have a gate length increased by 16nm) is 13% lower than the CS delay generator, but its maximum delay is 84% higher, resulting in an global energy overhead due to buffers insertion for fixing the hold time constraint.

One can be surprised by the fact that a higher gate length leads to a higher delay dispersion. However, these results are not in contradiction to the design theory. What it is well known is that the ratio $\sigma/\mu$ decreases with the number of gates or the increase of transistor dimensions. But, we have to keep in mind that while $\mu$ increases, $\sigma$ increases as well, as illustrated on figures 3.3 and 3.7.

Figure 3.8 compares the CS delay generator and the conventional one with different architectures: the replica DG proposed in [51], one, two or three half Schmitt triggers, and five or seven stages. As explained in section 3.1, adding inverters leads to a tradeoff between energy and robustness. Half Schmitt triggers

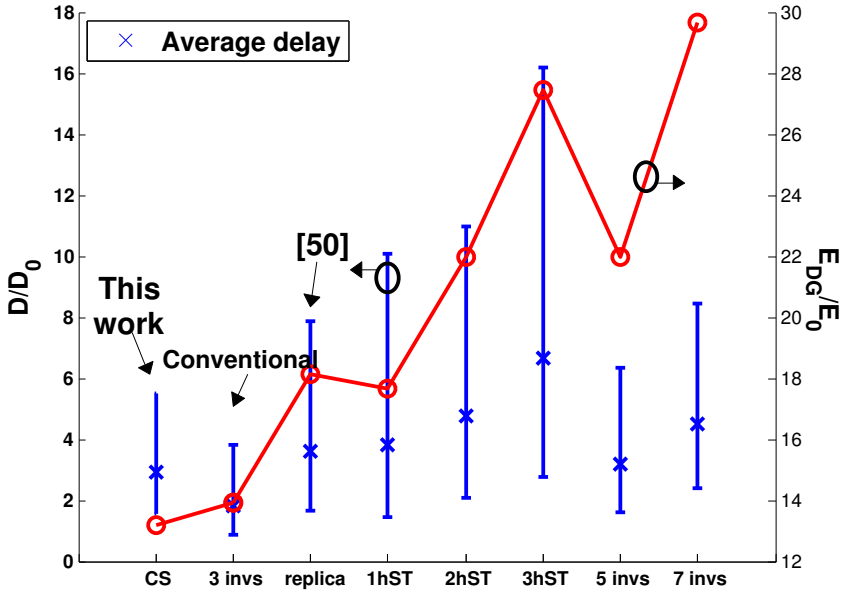



**Figure 3.8:** Comparison of energy and delay performances in subthreshold regime ($V_{dd} = 0.3V$) for difference DG architectures.

present a lower robustness than the chain of seven inverters, with a higher energy dissipation due to the hold time penalty. The replica DG decreases both the energy and hold time penalty compared to half Schmitt triggers and inverters chain. However, the current-starved DG exhibits, for the same robustness, a lower energy consumption and delay spread. Finally, compared to the conventional three inverters chain, the hold time penalty of the CS is 50% higher and its energy consumption is 6% inferior.

In summary, the CS pulse generator and the conventional are the best candidates for energy-efficient circuits. The conventional pulse generator provides the lowest global energy consumption but its minimum delay is lower than the CS delay generator, leading to a weaker robustness. Therefore, in order to quantify the robustness face to environmental variations, the next section studies the rate of failure of complete pulsed-FF architectures, one with the conventional DG and the other with our proposed current-starved DG, in post-layout simulations.

### 3.3.1 Layout comparison

To show the performance of our proposed delay generator, layouts of complete TGPL-Data flip-flops have been made (figure 3.9). The flip-flop has a minimum driving strength in the technology library, and is designed for reaching a minimum energy-delay product. The first layout contains a conventional delay

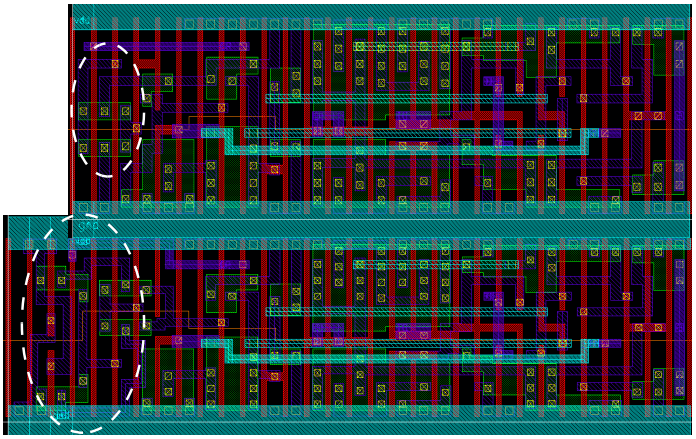


**Figure 3.9:** Layout of TGPL-Data (mininmum EDP and driving strength) with the conventional DG (above) and the current-starved-like DG (below). An overhead of three fingers is paid for an impressive increase of robustness.

generator, and the second contains the current-starved delay generator with the selected sizing ($W_{CS} = 150nm$ and $L_{CS} = 38nm$).

Let us notice that every transistor, in the PG and the D-to-Q path of the pulsed-FF, has the same threshold voltage $V_{th}$. One could argue that if the transistor of the DG had a higher $V_{th}$ than the transistors in D-to-Q path, the drawback of minimum delay of the conventional DG would be resolved. However, at the same subthreshold supply voltage, the delay dispersion between two $V_{th}$ is highly significant [75]. Thus, guaranteeing the minimum delay will again result in high energy overhead, even for the conventional DG.

Figure 3.9 shows that three additional fingers are needed to make to complete layout with our proposed pulse generator. To give a comparison, the TGPL architecture of minimum drive is composed of 29 fingers and this number obviously increases for higher drives, while the size of the DG does not. The robustness of these flip-flops is tested by writing logical 1 and 0 in worst environmental conditions. Table 3.1 shows the results of a 1000 runs Monte-Carlo simulation for several PVT conditions. As we see, the number of latching failure is dramatically reduced.

The next section presents silicon measurements which were performed to definitely show the gain of robustness provided by our current-starved architecture.

## 3.4 MEASUREMENTS

A test chip has been fabricated in 28nm FDSOI technology to test the functionality of a single FF for the four following architectures:

**Table 3.1:** Robustness comparison with 1000 Monte-Carlo runs of post-layout simulations in subthreshold regime (0.3 V)

| PVT conditions | # Failed | |
|---|---|---|
| | Conventional | Proposed |
| FS -40° | 14 | 0 |
| FS 85° | 3 | 0 |
| SF -40° | 60 | 1 |
| SF 85° | 29 | 0 |

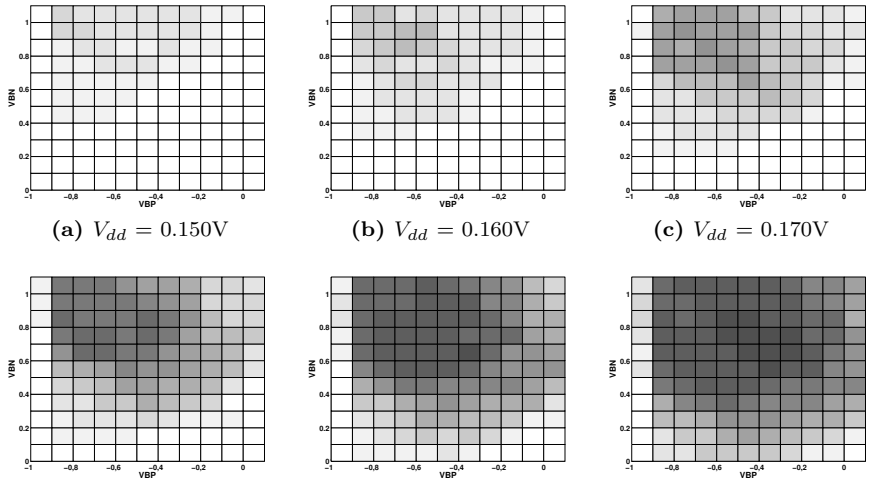
- a conventional $C^2MOS$ master-slave topology from a industrial standard-cell library
- the TGPL-Data architecture with conventional delay generator
- the TGPL-Data architecture with our proposed current-starved delay generator
- the TGPL-Clk architecture with our proposed current-starved delay generator

All these flip-flops shared the same clock, input data, PMOS back bias (vdds), and NMOS back bias (gnds) signals and have been designed in LVT feature. This section presents firstly the yield of the FFs in the (vdds,gnds) space. Because of the LVT feature, only a forward back biasing (FBB) is shown. Afterwards, the shape of the yield in the (vdds,gnds) space is explained by theory in simulation. Finally, the average yield is compared for each $V_{dd}$ between the four different FFs. The robustness gain of our current-starved architecture is thus highlighted.

### 3.4.1 Yield in the (vdds,gnds) space

The functionality of a single FF, for a large set of ($V_{dd}$,vdds,gnds), has been tested at room temperature on 64 chips. This computed yield is showed in figure 3.10 (The darker is the square, the higher is the yield) for each point of the set (vdds,gnds) and some $V_{dd}$ values.

We clearly see that at each $V_{dd}$, there is(are) one or several couple(s) (vdds,gnds) providing a maximum yield. The number of optimal couple(s) and the average yield extremely rapidly increase relatively to the small increase of $V_{dd}$ between each figure (10mV). From that/those optimal couple(s), the yield gradually decreases until 0%. However, we see that the yield gradient is not the same in every direction. The yield is actually maximum along a 45° line passing through the optimal points and this line is bounded on the four corners direction. Thus, the yield has approximately the shape represented in figure 3.11 which will be explained in the following section.




**(a)** $V_{dd} = 0.150$V **(b)** $V_{dd} = 0.160$V **(c)** $V_{dd} = 0.170$V

**(d)** $V_{dd} = 0.180$V **(e)** $V_{dd} = 0.190$V **(f)** $V_{dd} = 0.200$V

**Figure 3.10:** Computed yield of TGPL-Data in the (vdds,gnds) space (25°C). The darker is the square, the higher is the yield.

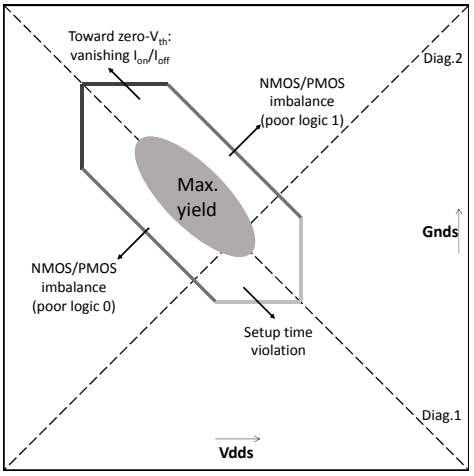



**Figure 3.11:** Schematic behaviour of the yield in the vdds-gnds space.

### 3.4.2 Analysis of the trend

In nanometer CMOS technologies, the functional yield, taking into account the output logic levels at ULV, was shown in [76] to be directly related to the ratio between $I_{on}$ in linear mode (low $V_{ds}$) and $I_{off}$ in saturation mode ($V_{ds} \approx V_{dd}$),

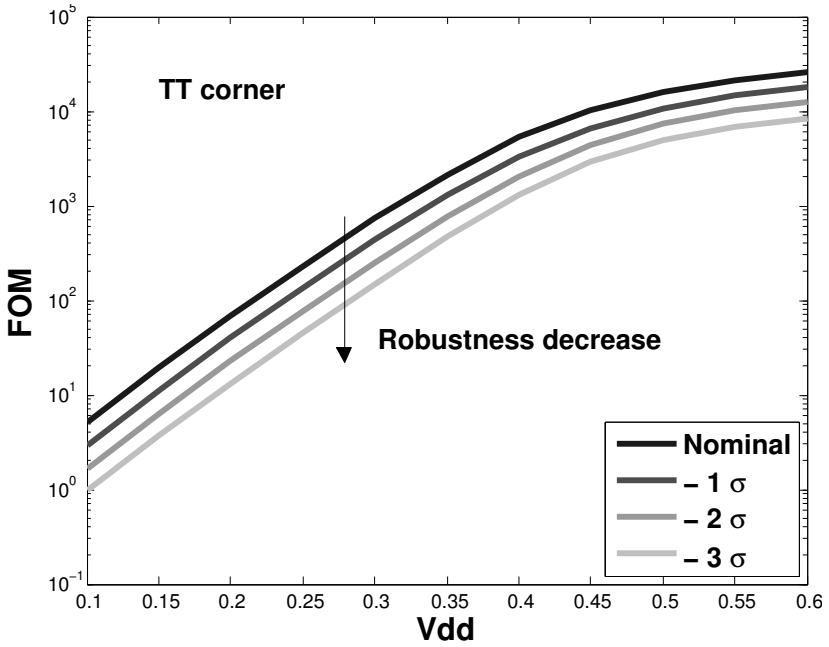



**Figure 3.12:** At ULV, the degradation of the FOM ($\frac{I_{on}}{I_{off}}$ ratio) leads to a degraded output logic level.

which is strongly affected by three parameters: the subthreshold slope, the DIBL effect, and the local variability (mismatch) [76]. Moreover, NMOS/PMOS imbalance also significantly affects the output logic levels [77]. To capture this trend, we define a new figure of merit (FOM) for functional yield with respect to output logic levels as:

$$FOM_{logic0} = \frac{I_{on,lin,N}}{I_{off,sat,P}} \tag{3.1}$$

$$FOM_{logic1} = \frac{I_{on,lin,P}}{I_{off,sat,N}} \tag{3.2}$$

$$FOM = min(FOM_{logic0}, FOM_{logic1}) \tag{3.3}$$

where

$$I_{on,lin} \longrightarrow |V_{GS}| = V_{dd} \quad ; \quad |V_{DS}| = 0.2V_{dd} \tag{3.4}$$

$$I_{off,sat} \longrightarrow |V_{GS}| = 0 \quad ; \quad |V_{DS}| = 0.8V_{dd} \tag{3.5}$$

The yield dependence with $V_{dd}$ is a well known behaviour which can easily be explained considering the evolution of the FOM with Monte-Carlo simulations (see figure 3.12).

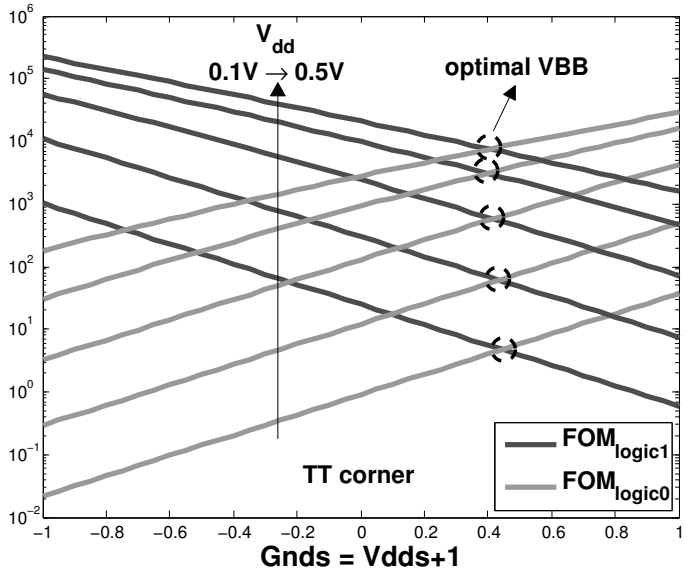



**Figure 3.13:** For Vdds and Gnds values evolving in the same way, an optimal point is found near the middle of the diagonal 2 (Fig. 3.11).

The two gray sides in figure 3.11 are explained by the imbalance between the PMOS and NMOS strengths, also called imbalance factor in [77]. Along diagonal 2 and its parallels, the threshold voltages evolve in opposite direction and so do the two components of the FOM, as illustrated in figure 3.13. Notice that the optimal point slightly varies with $V_{dd}$.

The dark gray boundary in figure 3.11 is also explained by the decrease of the FOM but for another reason. Along diagonal 1, vdds and gnds are symmetrical and thus the currents evolve in the same direction. Nevertheless, continuing to decrease the threshold voltage might eventually lead to a zero-$V_{th}$ transistor. That is why the current ratio progressively decreases when we go to strong FBB as depicted in figure 3.14. As a result, FBB substantially increases the speed of the circuit at ULV (see Chapter 2) but this performance gain is at the expense of lower robustness. *Medèn ágan.*

Finally, the light gray boundary in figure 3.11 is due to setup time violation. The clock period of our test pattern was 2000ns and, at ULV with nominal threshold voltages, FFs are too slow to switch before the triggering clock edge.

Let us notice that, theoretically, symmetrical RBB does decrease the speed but does not decrease the robustness. Indeed, in sub-threshold regime, the on- and off- currents of the PMOS and NMOS follow the same slope and thus evolve in the same way. However, in the 28nm FDSOI technology, the sub-threshold slopes are not the same for PMOS and NMOS. If their $V_{th}$ is increased by the same amount, the on- and off-currents of the PMOS decrease more slowly than

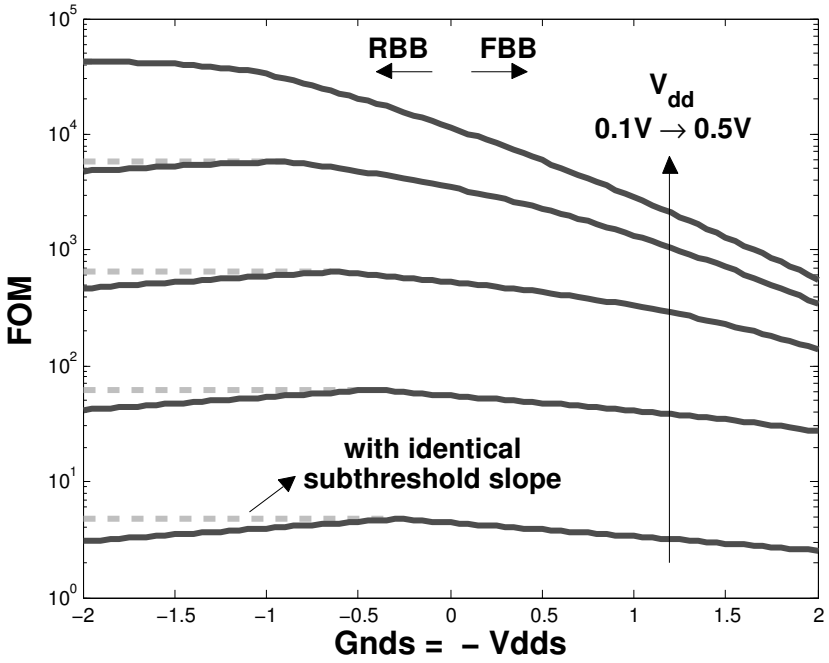



**Figure 3.14:** Strong FBB increases the speed but decreases the robustness. In 28nm FDSOI technology, the non-identical subthreshold slopes cause a robustness decrease also for RBB.

the ones of the NMOS. Thus the ratio $\frac{I_{on,lin,P}}{I_{off,sat,N}}$ progressively decreases in our 28nm FDSOI technology, which explains the change of behaviour in the RBB region in figure 3.14. Therefore, the light gray boundary is explained by both timing violation and logic level failure in our case.

### 3.4.3 Yield comparison

The figure 3.15 compares the average of the yield in the (vdds,gnds) space in terms of the supply voltage.

If the average yield at a certain $V_{dd}$ is higher for an architecture than another one, it means that this architecture is more robust to a back biasing variation, thus a $V_{th}$ shift. Similarly, if the same average yield is reached at a lower $V_{dd}$, one architecture is more robust to a variation of supply voltage than the other one. We can see in figure 3.15 that the lack of robustness of the explicit pulsed-FF topology, compared to Master-Slave topology, is now completely filled thanks to the current-starved technique.

Below 0.3V, the TGPL-Clk presents a robustness slightly higher than the other pulse-triggered architectures. In addition to its current-starved delay gene-

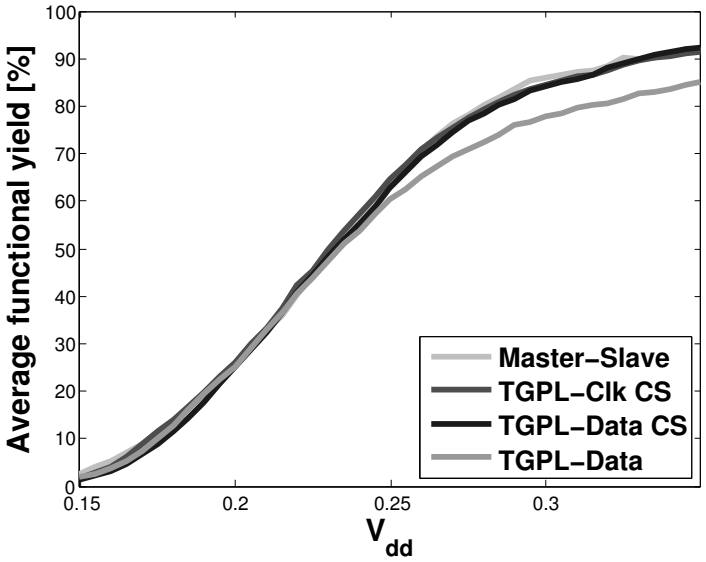



**Figure 3.15:** Yield average over the whole (vdds,gnds) space. The explicit pulsed-FFs presents, now, the same robustness as Master-Slave topology.

rator, its D-to-Q is much faster at ULV thanks to its lower stack in the input inverter. Thereby even if the generated delay is short, the FF has a higher probability to latch the input data.

The last two architectures help us to compare the yield with and without the current-starved delay generator discussed previously. As we see in figure 3.16, the current-starved DG provides the same yield value at a supply voltage up to 45mV inferior and a yield 7.5% superior at the same $V_{dd}$. It means that, for the same yield, FFs can work in ULV operations at a $V_{dd}$ 45mV lower than without the current-starved technique.

Moreover, we see that the robustness gap between TGPL-Data without CS technique and the master-slave FF is eliminated thanks to the current-starved DG. As a reminder, the lower robustness was the main disadvantage of pulse-triggered FF at ultra-low voltage. From our results coming from the measurement of 63 FFs, we can affirm that our technique has a significant impact on the pulsed-FF robustness.

After the current-starved technique in the delay generator, the next and final section shows a second way to increase the energy-efficiency of pulsed-FFs: the implementation of the additional functionalities in the pulse generator.

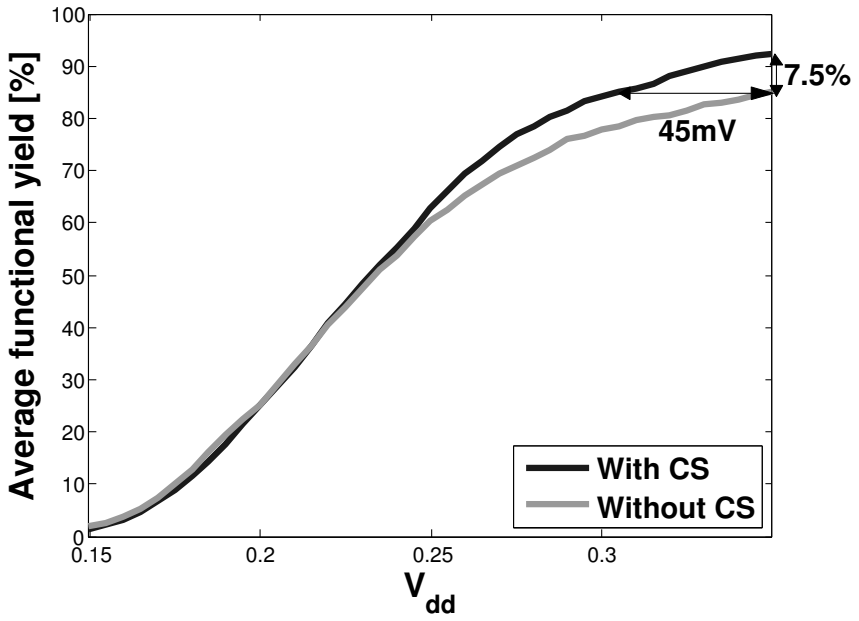



**Figure 3.16:** The current-starved (CS) delay generator provides a notable gain for the yield.

## 3.5 ADDITIONAL FUNCTIONALITIES

As already mentioned, flip-flops present hardly nothing but the three basic connections (D, CLK, Q) in an industrial standard-cell library. In addition to the scan and reset functionalities discussed in the previous chapter, the set and enable functions are also familiar in a common library. The set function is exactly the same as the reset excepted that the output presents a high value. Thus, the transistors used to perform the set function are simply the complementary of the transistors in the reset version. The enable function forbids the writing of the flip-flops without modifying the output. In every case, the writing system must be disabled to avoid writing two different data.

In master-slave flip-flops, these functionalities are implemented inside the two latches, by modifying the (tristate) inverters composing them. Disabling the writing system cannot be implemented in the clocked inverters providing the CLK and $\overline{CLK}$ signals because if the signal becomes inactive during the high level of the clock, a triggering edge will occur despite the global triggering edge is passed. Therefore, the flip-flop will not be functional any more.

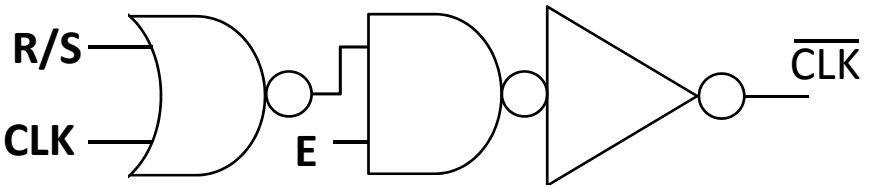



**Figure 3.17:** (Re)Set and Enable functionalities are easily implemented in the pulse generator.

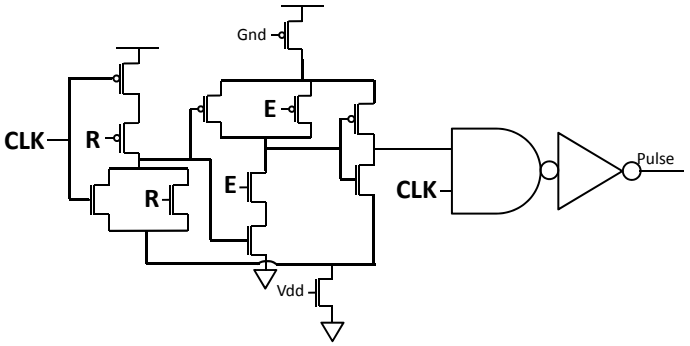



**Figure 3.18:** (Re)Set and Enable current-starved pulse generator.

As the latch of a pulsed-FF does not work on the levels of the clock but on a edge, the clock system can be easily modified without disturbing the correct functionality of the FF. As we can see in figure 3.17, the pulse generation can be disabled with a NAND and/or NOR gate in the delay generator. It means that only two additional transistors are needed to perform each additional functionality, plus two transistors in the latch for the set and reset functions.

Obviously, the functionalities can be performed with the signal or its complementary, depending on layout or design considerations. After the preceding sections, we are tempted to use the stacking of NAND and NOR as native current-starved technique. Unfortunately, the stack may not be on the delay path. Indeed, when the functionality signal is active, the $\overline{CLK}_d$ signal must be pulled to ground whatever the value of the clock is. Thus, the active signal is on the PMOS of NAND gates and NMOS of NOR gates. But, during normal operations, the $\overline{CLK}_d$ signal is also pulled to zero, therefore using the same pulling system in NAND and NOR gates. Hence, the use of the current-starved technique should be modified as represented in figure 3.18.

In master-slave topology, the additional transistors laying in the latches directly impact the timing and/or the energy performances. Indeed, as adding transistors will increase the stack, either the current is lowered or the gate width

increases and so does the energy consumption. In pulsed-FFs, as the additional transistors lay in the pulse generator but are not directly on the delay path, the impact on the delay and energy performances is significantly lowered.

**Table 3.2:** Area [$\mu m^2$] comparison between a conventional master-slave and the TGPL-Data architecture

| Topology | Original | Resettable | Resettable and Enable |
|---|---|---|---|
| MS | 3.75 | 4.4 (+17%) | 5.38 (+43%) |
| pulsed-FF | 5.38 | 5.71 (+6%) | 5.87 (+9%) |

Layouts of scannable, resettable and/or enabling TGPL-Data flip-flops have been made for the comparison. Table 3.2 compares the area of master-slave FFs from industrial library and pulsed-FFs with the same amount of functionalities. We see that, despite the efficient implementation of reset and enable functions in the pulse generator, the master-slave architectures exhibit a lower area than the pulse-triggered flip-flops for the same functionalities. Nevertheless, Table 3.3 compares the performances of these same FFs. As we can see, the impact of the additional functionalities is significantly much lower for pulsed-FFs than master-slave which shows that they can be efficiently integrated in a complete standard-cells library. Moreover, a simple calculation shows that the complete scannable, resettable and enabling master-slave has a area-energy-delay product (AEDP) almost 3.5 times superior to the AEDP of the corresponding pulsed-FF. The gain in energy-efficiency is now clearly highlighted for UWVR and ULP circuits.

**Table 3.3:** EDP [ps.fJ] comparison between a conventional master-slave and the TGPL-Data architecture (post-layout simulations)

| Topology | Original | Resettable | Resettable and Enable |
|---|---|---|---|
| MS | 580 | 721 (+24%) | 953 (+64%) |
| pulsed-FF | 236 | 249 (+5.5%) | 259 (+9.7%) |

## 3.6 CONCLUSION

In this chapter, a new pulse generator for ultra-wide voltage range pulse-triggered flip-flops is presented. First of all, the key issues of pulse-triggered flip-flops at ultra-low voltage (ULV) were presented. It was explained that the minimum

generated delay must be large enough in order to ensure the correct functionality of the pulsed-FF and, on the other hand, the maximum generated delay should be as small as possible to minimize energy overhead. The proposed DG consists of using current-starved-like inverters in the delay chain, with a PMOS and a NMOS always in on-state. This architecture allows a great flexibility in design, by sizing the current-starving transistors without impacting the dynamic energy. Several architectures of delay generators have been compared with one chosen sizing of our delay generator. All of them present at least one drawback in the figures of merit characterizing delay generators: minimum delay, delay dispersion and energy consumption. Post-layout simulations have been performed to compare the robustness of our pulse generator with the conventional one. It is shown that, for an area penalty of only three fingers, the number of latching failures at ultra-low-voltage is dramatically reduced.

Afterwards, silicon measurements were presented to study the robustness improvement of the current-starved DG. Moreover, we showed that, based on our results, the robustness gap between pulsed-FFs and master-slave structures is compensated thanks to the current-starved technique.

Finally, the implementation of additional functionalities in pulse-triggered flip-flops was studied. In master-slave (MS) topology, the additional transistors needed to carry out these functionalities, lead to a large increase in area and energy-delay product (EDP). On the other hand, it has been shown that, for our pulsed-FFs, the reset and enable functions can easily be performed and implemented in the pulse generator. The EDP overhead is only 9% for pulsed-FFs and 64% for master-slave, while the AEDP of the biggest pulsed-FF is 3.5 times smaller than the biggest MS.

CHAPTER 4

# INTEGRATION AT BLOCK LEVEL: SYNTHESIS AND PLACE&ROUTE CONSIDERATIONS

## Abstract

This chapter focuses on the integration of our energy-efficient explicit pulse-triggered flip-flops in large digital circuits, *i.e.* at block level.

First, the phenomena of clock skew encountered in the design at block level is briefly discussed and its impact on our explicit pulse-triggered flip-flop architectures is exposed.

Then, the design of a conditional capture pulsed-FF is established with the same methodology as in Chapter 2. Several variants are compared in the energy-delay ($E - D$) domain in order to select the most energy-efficient. This conditional capture technique disables the pulse generator when there is no data activity and thanks to that, this explicit pulse-triggered architecture exhibits a lower energy consumption than the master-slave architecture available in an industrial standard-cells library. Then, it is shown that this new architecture gives another degree of freedom in the energy-delay-area tradeoff faced by the automatic synthesis tools. Using the lowest FF energy consumption while meeting the timing constraint for each path, provides a Pareto-optimum-like energy-efficient circuit synthesis [1].

Afterwards, the sharing of the pulse generator (PG) of explicit pulse-triggered flip-flops is studied. The pulse generator represents the largest part of the energy consumption of pulsed-FFs, but this part can be divided by the number of latch when it is shared. We show that, after a given number of latches sharing the same PG, the energy and area per FF are actually lower than the energy dissipation and area of a master-slave FF. To conclude, a complete 16x32bits register file was laid out using the energy-efficient scannable $C^2$MOS-Data latch of Chapter 2, the enable and resettable current-starved PG of Chapter 3 and the PG sharing property. It is shown that, compared to a master-slaved based register file, our energy-efficient explicit pulse-triggered flip-flops provide lower energy consumption (-10%) above a supply voltage of 0.6V and a lower area (-14%).


## Contents

## 4.1 INTRODUCTION

So far, we have clearly seen that explicit pulse-triggered flip-flops (pulsed-FFs) are much faster than master-slave flip-flops whereas they suffer from energy and area penalties. After guaranteeing the robustness of pulsed-FFs at gate level in Chapter 3, this chapter studies several techniques to decrease their energy, area or both at block level and finally reaches better performances than master-slave (MS) flip-flops.

In Section 4.2, the impact on the clock skew on the explicit pulse-triggered flip-flop architecture is discussed. We show that soft-edge property allows a certain immunity and facilitates the timing closure during clock tree synthesis.

In Section 4.3, a conditional capture technique is applied and studied with the aim of saving energy. Then, several architectures are compared in the energy-delay domain, as in Chapter 2. It is shown that this architecture exhibits a lower energy consumption but also lower speed than the MS, $C^2$MOS and TGPL architectures developed in Chapter 2. Thereby, it provides another design point in the energy-delay ($E - D$) tradeoff of flip-flop design and, with the help of master-slave topology, may help the tools to produce a Pareto-optimum-like netlist in an energy point of view.

As mentioned in Section 1.4, one of the advantages of the explicit pulse-triggered flip-flop is the shareability of its pulse generator. This is one of the ideas proposed in the literature to reduce the FF energy budget. In Section 4.4, the energy consumption per flip-flop is studied for shared pulse generators and compared to the energy consumption of the master-slave flip-flop. Following this idea, a complete register file with the energy-efficient scannable $C^2$MOS-Data latch and a shared current-starved pulse generator with the enable and reset functionalities, is developed in Section 4.5. By comparing to register files based on master-slave flip-flops, it is shown that the explicit pulsed-FF finally presents a lower energy consumption, even without the conditional capture, and a lower area than master-slave thanks to pulse generator sharing. Finally, Section 4.6 concludes this chapter.

## 4.2 CLOCK SKEW ABSORPTION

After the clock tree synthesis, it appears that the clock signal do not exactly arrive at the same time at the leaf level. Even in neglecting the local variations, a static timing analysis of the clock tree shows a certain difference between the clock signal arrivals. If a flip-flop receives its clock signal a time $\Delta T$ before the preceding flip-flop in a path, the data has $\Delta T$ time less to pass through the combinational logic and reach this flip-flop. This difference is called the clock skew and it directly penalises the speed of the circuit. On the other hand, if designers set a too small clock skew, the clock tree synthesis may become very difficult and the use of more and/or larger clock buffers might become necessary, leading to additional energy consumption.

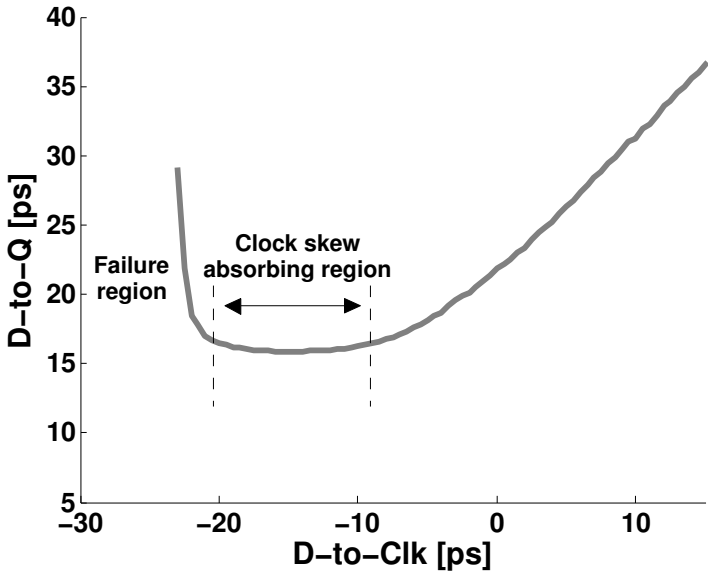



**Figure 4.1:** Timing regions and characteristics for TGPL flip-flop (nominal voltage.

More precisely, clock skew is modelled as a window around the nominal arrival time where the actual clock transition may occur [78]. A change in the D-to-Clk delay might cause a fluctuation in the effective input-to-output delay. Nevertheless, the soft-edge property of the pulse-triggered flip-flops leads to a D-to-Q behaviours as shown in figure 4.1. As we can see, over a large window, the D-to-Q delay does not vary that much with the D-to-Clk delay, thus the clock arrival time. That allows to tolerate a larger clock skew in the clock tree without modifying the timing performances. Thus, the constraints on the clock tree synthesis are lower and lead to a save in energy.

In conclusion, the pulse-triggered flip-flops present a large clock skew absorption, which tends to facilitate the timing closure and improve the timing and energy performances of the circuit.

## 4.3 CONDITIONAL CAPTURE ARCHITECTURE

We have seen that, in our 28nm FDSOI technology, the most energy-efficient pulsed-FF architecture, over a wide range of supply voltage and driving strength, is the $C^2$MOS-Data architecture, with the current-starved delay generator (DG). It is an explicit pulse-triggered topology, where the distribution of energy is

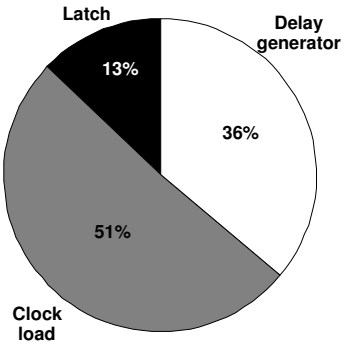


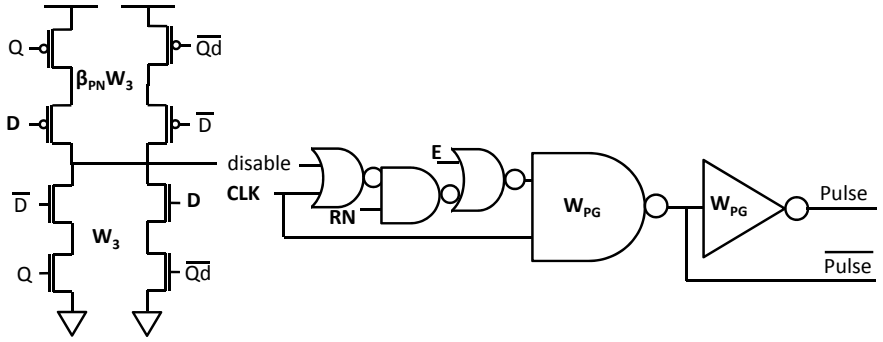

**Figure 4.2:** Repartition of the energy dissipation in C$^2$MOS architecture (minimum EDP sizing). Corner TT, nominal voltage (1V), temperature = 70°C.

represented in figure 4.2 for minimum driving strength. We clearly see that the main source of energy consumption is the pulse generator (PG), made of the delay generator and the gates controlling the clocked transistor of the latch (included in clock load). The part of the delay generator is less predominant with a higher driving strength, because the energy consumption of the latch is higher. However, as mentioned in Section 2.3, the lowest driving strength is the most used in the synthesis of industrial low-power microprocessors. Therefore, as we focus the energy-efficiency, only the energy consumption of the lowest driving strength is of primary importance. This section studies one of the ways to tackle the high energy consumption of the pulse generator based on the "xored" input-output technique. This idea is to compare the current input with the current output and disable the FF latching mechanism if they are identical [80]. Firstly, we show how to implement the XOR technique in our efficient pulse generator with a brief discussion about the position of the pseudo-XOR gate. Secondly, we present various implementations in the latch and compare them in the energy-delay ($E - D$) domain with the same methodology of Chapter 2.

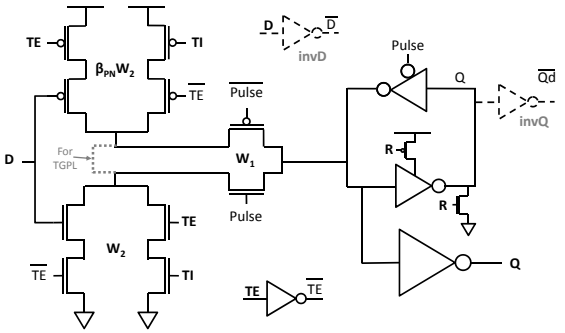
### 4.3.1 Pseudo-XOR gate in explicit pulsed-FFs

The five stages of the PG (3 minimum-sized current-starved inverters, a NAND gate and an inverter) are activated at each clock cycle, even without data activity. In addition, the clocked transistors of the latch are also charged and discharged even when the input has not changed. To tackle this useless energy consumption, the XOR-architecture represented in figure 4.3 is proposed with a slightly different approach from [80]. Using the explicit pulse property, a pseudo-XOR gate is added in order to disable the pulse generation when both the input (D) and the output (Q) are the same, *i.e.* when latching is not needed.

Let us notice that this technique cannot be applied with edge-triggered flip-flops, like the master-slave topology. Indeed, in order to disable the latching,




**(a)** The conditional pulse generator (current-starved circuitry not shown for clarity).




**(b)** The C$^2$MOS-XOR latch with additional inverters for $\overline{D}$ and $\overline{Q_d}$ signals.

**Figure 4.3:** Schematic of the TGPL/C$^2$MOS-invQ/invD architectures.

the conditional capture technique must maintain a closed master latch and open slave latch during the triggering edge of the clock. But if the incoming data changes between the triggering and the non-triggering clock edge, the conditional capture is being disabled and the FF will see a new clock level. This will create a triggering edge on the FF *after* the global triggering edge of the circuit, thus leading to functional failure. To sum up, it is impossible for the edge-triggered flip-flop to make the difference, excepted by exhibiting a hold time of half $T_{clk}$ which is obviously non-acceptable in energy-efficient circuits.

The position of the pseudo-XOR output in the delay generator will impact the setup time and the hold time of the flip-flop, and the number of stages which flip at each clock cycle. If the XOR gate is close to the NAND gate of the PG, the input data edge has more time to reach the FF input before the end of the propagation of the triggering clock edge. Therefore, the setup time and the hold

time evolve in the same direction. Since $T_{setup} - T_{hold}$ is the key parameter for the useful-skew technique ([55, 56] Section 1.4), the only criteria for the position of the NAND is the energy dissipation of the inverter chain. Obviously, there is less energy overhead if the first inverter is disabled when D is identical to Q (figure 4.3). Finally, the position of the enable (E) and reset (R) signals (see Section 3.5) is a design choice depending on the use rate of those signals in the application.

Afterwards, let us point out that the setup and hold times are now positive and negative, respectively. Indeed, the incoming data must now be valid sufficiently before the triggering clock edge so that the *disable* signal enables the pulse generator, *i.e.* the $\overline{CLK}$ signal has a high logic level. Similarly, the hold time becomes negative because if an input data comes after the triggering clock edge, the NAND gate output in the pulse generator is already pulled to one.

### 4.3.2 Architectures comparison

A two-inputs CMOS XOR gate needs the two inputs plus their complementary signals. $D$ and $Q$ signals can be easily taken from the input signal and the bistable element node not laying on the D-to-Q path, without degrading the timing performances (assuming identical input slope for $D$). The $\overline{D}$ and $\overline{Q_d}$ (for $\overline{Q}$ delayed) signals can be taken directly in the D-to-Q path, or by adding inverters connected to the pseudo-XOR gate. Notice that providing $\overline{D}$ signal in the D-to-Q path means to switch to a TGPL architecture. Figure 4.4 shows a comparison performed on the different XOR configurations in the $E - D$ space, where the sizes of the transistors in the XOR gate are additional sizing variables.

The lowest energy consumption is still provided by C$^2$MOS architectures, because of the lowest junction capacitance (dis)charged in the input tri-state inverter (see Chapter 2). Again, the gap is small because the junction capacitance of FDSOI technology is highly lowered compared to bulk. The minimum energy-delay product (EDP) is reached by the C$^2$MOS topology with an additional inverter for $Q$ (C$^2$MOS-invQ), as represented in figure 4.3, for a D-to-Q delay of approximately $6D_0$. Nevertheless, we can see that directly after this small window where C$^2$MOS-invQ is the most energy-efficient, the TGPL-invQ-D architecture (with an additional inverter for $Q$ but with $\overline{D}$ taken directly from the input inverter) becomes the most energy-efficient for the high-speed figures of merit. However, as explained in Section 2.3.4, the back biasing technique allows designers to cover the $E - D$ domain with the same architecture and even more efficiently than with the sizing methodology. Therefore, to save design time, designers can choose to use only the TGPL topology, with a reverse body bias when performances are not under consideration.

Finally, the C$^2$MOS-XOR architecture with TGPL-Clk, C$^2$MOS-Data and the master-slave C$^2$MOS FF are compared in the $E - D$ domain (figure 4.5).

First of all, we see that our conditional capture technique, disabling the pulse generator when $D$ and $Q$ are identical, provides a lower energy consumption

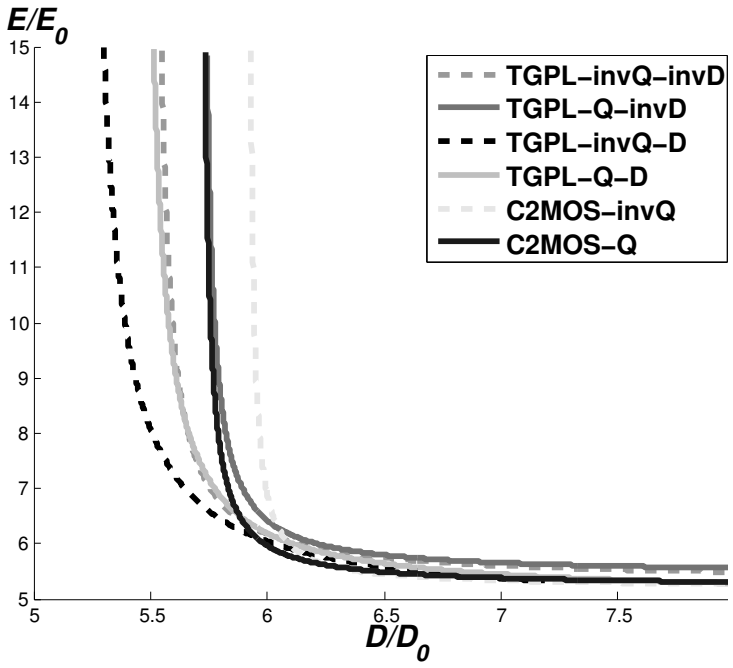



**Figure 4.4:** Energy-Delay with several different XOR ($V_{dd}$ = 1V, driving strength X1 , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, temperature 70°C).

than the master-slave architecture. Consequently, ULP circuits, targeting a very low energy consumption without hard speed constraints, might use this explicit pulsed-FF architecture. Moreover, let us remind that the problem of positive hold time is alleviated because the conditional XOR-technique induces a positive setup time and negative hold time.

**Table 4.1:** Area [$\mu m^2$] comparison between the four most energy-efficient FF architectures.

| mC$^2$MOS MS$^\dagger$ | C$^2$MOS-Data | TGPL-Clk | C$^2$MOS-Xor |
|---|---|---|---|
| 4.4 | 5.4 | 6.7 | 6.7 |

† Highly optimised layout from industrial library

Secondly, we can imagine a Pareto-optimum-like curve is provided in the $E-D$ space by mixing different FF architectures, at the cost of area overhead (see Table 4.1). It means that the synthesis tools have the opportunity to choose the most energy-efficient FF for a path, depending of the timing constraints of this path,

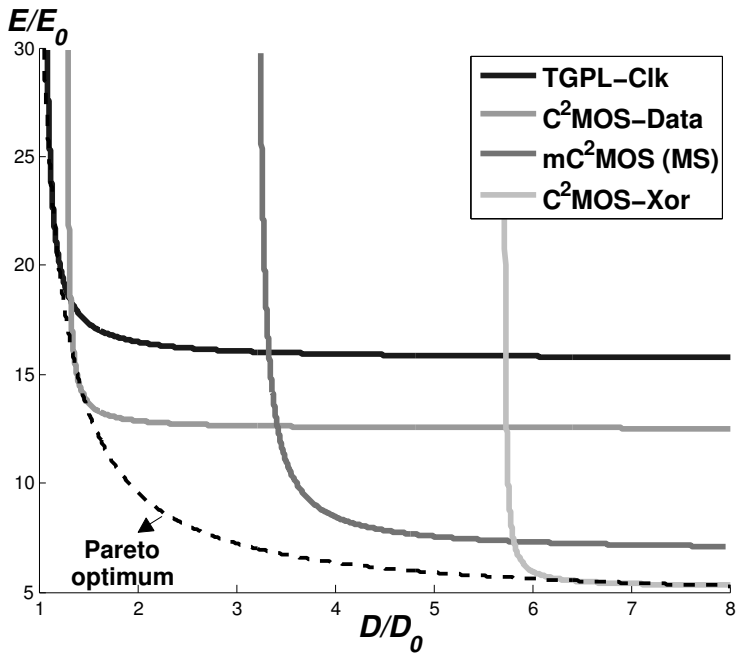



**Figure 4.5:** EECs for the four interesting architectures at $V_{dd} = 1$V (driving strength X1 , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, temperature 70°C). The dashed curve is only there to illustrate the Pareto optimum idea.

in order to provide a minimum energy consumption. This pareto optimum point in synthesis is extremely interesting for UWVR circuits where both speed and power are under consideration.

## 4.4 PULSE GENERATOR SHARING

Sharing the pulse generator has been studied many times from an energy-efficiency point of view [81, 82, 83, 84]. The objective of sharing one pulse generator with N latches is to get a lower energy consumption than using 2N latches with master-slave topology. Most of the papers in the literature work in the superthreshold regime where the random variations can be more easily handled to ensure a correct and sufficient slope of the pulse signal after the Place&Route steps. In near- and subthreshold regime, we consider that the variability of the pulse signal is too high (see Chapter 3) to deal with a variable PG output capacitance. Therefore, we recommend a single block of many latches and one pulse generator which can be characterized independently. Let us notice that a block of FFs can be useful in many applications: synthesised register files or any other

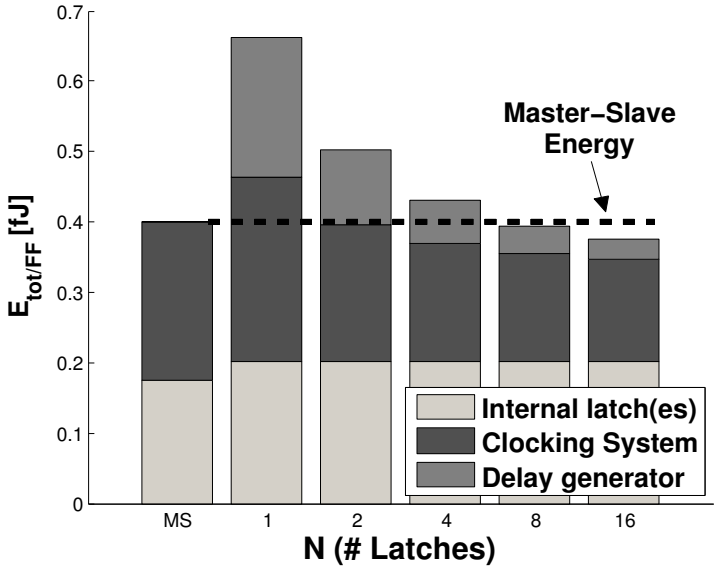



**Figure 4.6:** The energy per FF with the number of latches sharing a PG. Bars represent the proportions of the energy consumed in the delay generator, the clocked part ($CLK$ and $Pulse$ signals), and the latch alone. $V_{dd}$=1V, minimum driving strength (X1), $T_{clk}$ = 40FO4, $\alpha_{sw}$ = 15%.

standard-cells based memory [85], or pipeline registers in low-power datapaths like the 16bits or 32bits ultra-low power CPU ([57, 86] respectively).

Figure 4.6 shows the energy consumption of a block of N latches sharing a PG compared to the energy consumption of one master-slave FF from an industrial standard-cells library (dotted line). The size of the transistor of the NAND and INV gates are adapted to reach the same slope for the $Pulse$ and $\overline{Pulse}$ signals. As we see, one PG shared with 4 latches dissipates almost the same energy per FF than the master-slave architecture. Then from 8 or more latches, shared pulse-triggered FF architecture is more energy-efficient than master-slave FF. Figure 4.6 also shows the relative proportion of the three contributions of the total energy of the bloc: the delay generator, the pulse generation and the external clock signal load (both included in the Clk load component) and the latch itself. When the pulse generator, meaning the delay generator plus the Clk load, is connected to only one latch, we can see that it represents more than 70% of the total energy consumption. This percentage clearly shows that it is the key point to be optimized in order to reach an energy-efficiency as the latch mainly deals

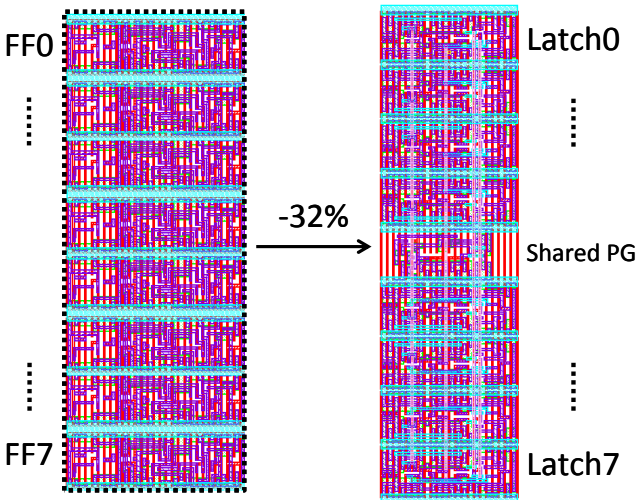



**Figure 4.7:** Pulse-triggered FF architecture shows a lower area per FF than master-slave. The block of 8 pulsed-FF (right) has 32% less area than 8 MS FFs, at the expense of the M3 layer utilization.

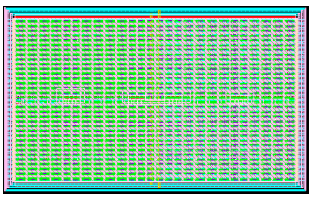
with the timing performances (and also robustness) of the FF. Proportionally to the total energy per FF, the Clk load component remains at the value ($\approx 40\%$) because the constant slope constraint is maintained. While the energy per FF due to the latch remains obviously roughly the same, the energy gain comes from the sharing of the delay generator. After 16 latches, the energy of the delay generator normalized by the number of FF becomes negligible and thus we see a stagnation of the benefit of sharing the pulse generator.

The area of a block of 8 latches sharing one PG is 32% lower than 8 times the area of a master-slave FF, considering the layout from an industrial library (figure 4.7). Therefore, after a certain number of latches, the area per FF is smaller for pulse-triggered topology with the drawback of using the first vertical layer (M3 in this technology) to spread the $Pulse$ and $\overline{Pulse}$ signals over the latches.

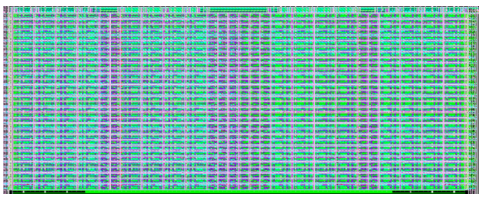
In the next section, we study a 16x32bits register file using the PG sharing developed in this section, in order to compare its timing, energy and area performances to master-slave topology based register file. It will be shown that the conclusion of this section remains the same when shared PG are used in a higher level application.

## 4.5 REGISTER FILE

In this section, we compare two register files: one with our pulsed-FFs and the other with conventional master-slave flip-flop. A register file is a small embedded



**(a)** Structured explicit pulsed-FFs based register file layout.



**(b)** Structured master-slave FFs based register file layout.

**Figure 4.8:** Layout comparison led to a 14% saving in area.

memory, synthesized and included directly into the logic. It is a key component in every von Neumann and Harvard microprocessor architecture and represents a non-negligible part of the microprocessor energy consumption [87]. A typical size for energy-efficient circuits is around 16 or 32 registers, each of them with the same number of bits as the datapath. For this small memory capacity, register files are more area- and power-efficient than SRAM and are faster in any case. Moreover, it is more robust (scalable in supply voltage) and can be easily integrated into the logic. Indeed, the fundamental bit cell of a register file is the flip-flop.

### 4.5.1 Design of register files

From all the studies performed in the previous chapters, we laid out a structured register file based on our energy-efficient explicit pulse-triggered flip-flop: a $C^2$MOS-Data architecture with minimum $E^1D^1$ product for the latch, a current-starved delay generator with enable and reset options, and one shared pulse generator for each register. In order to provide realistic conditions, we chose the same characteristics as the register file of the low-power dedicated Cortex-M0 microprocessor: 16 registers of 32 bits, one synchronous write port, and two asynchronous read ports. Each 32-bits register is composed of one pulse generator placed above and below 16 latches sharing the $Pulse$ and $\overline{Pulse}$ signals. The write port is implemented by a conventional 4-to-16 decoder, composed of NAND, NOR and inverter gates. Each read port is composed of two stages of 4-to-1 multiplexer for every bit. The clock network is an unbuffered H-tree 0 level since [57, 88] showed that this configuration provides lower slew and skew variations at ultra-low voltage than buffered clock-tree. A global reset and scan-enable pins are available, as well as 32 scan-inputs.

The pulse generator contains an enable signal $E$ connected to a decoder output. It allows to select the right register in the bank and, at the same time, it acts as a clock gating system. On the other hand, the MS-based register file contains latches at the decoder outputs performing clock gating.

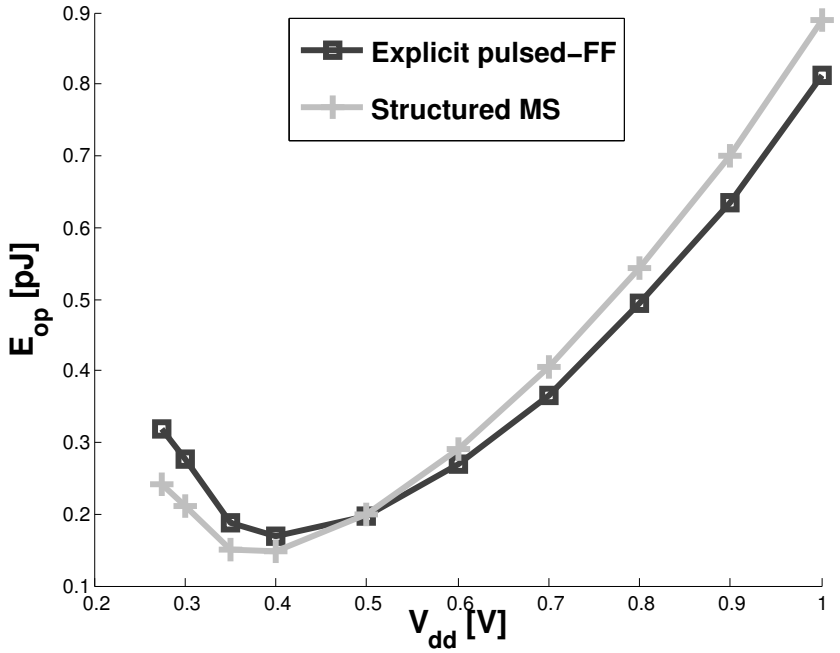



**Figure 4.9:** Energy per operation over a wide voltage range (Gnds = Vdds = 0V and $\alpha_{rate} = 15\%$). Explicit pulsed-FFs based register file presents a lower energy consumption for super-threshold $V_{dd}$ and an optimal one at 0.35V.

In the following section, the timing, area, and energy performances of the structured pulsed-FF-based and MS-based register files are compared.

### 4.5.2 Comparison of energy-delay-area performances

First of all, figure 4.8 shows the layouts of the pulsed-FF-based and master-slave-based register files. The area of the structured pulsed-FF-based register file is $40.8\mu m \times 72.8\mu m = 2970.4\mu m^2$ while the area of the MS-based register file is $40.8\mu m \times 83\mu m = 3386.2\mu m^2$, both including the multiplexers used for reading, the clock buffer, the write decoder and the clock gating system. Consequently, our pulsed-FF-based register file presents an area 14% lower than the MS-based one.

Figure 4.9 compares the average energy consumption per operation between the pulsed-FF-based and master-slave-based register files. These numbers are computed from SPICE simulations with the testbench presented in Annex C, where the clock period $T_{clk} = 40FO4$ is adapted for every supply voltage. For supply voltage higher than 0.5V, the average energy consumption per operation $E_{op}$ is lower for the pulsed-FF-based register file. Then, the higher leakage current

presented by the pulsed-FF leads to a lower $E_{op}$ for MS-based register file. Indeed, at low voltage, the leakage energy is more and more predominant and thus the leakage current penalizes the pulsed-based structure.

In parallel to this, a RTL code representing our register file operations has been synthesized, placed and routed by commercial tools. The RTL code performs register file operations with the same number of pins and ports. An industrial 28nm FDSOI LVT standard-cells library has been characterised at nominal voltage (1V) and ultra-low voltage (0.35V). Then, two synthesises have been performed with these two libraries. In both synthesises and Place&Route (P&R) in the two points of characterisation, the clock period has been set to reach the energy-efficient limit of the register file. Table 4.2 compares the area of the structured and automatic placed layouts, as well as the energy-delay performances for two supply voltage values.

**Table 4.2:** Energy-delay-area comparison of our explicit pulse-triggered flip-flop based and the master-slave (MS) based register files at $V_{dd} = 1V$ and $V_{dd} = 0.35V$.

| | Pulsed-FF | Master-slave structured | Master-slave 1V library | Master-slave 0.35V library |
|---|---|---|---|---|
| Area $[\mu m^2]$ | 2970.4 | 3386.2 (+14%) | 4123.5(+39%) | |
| $D_{read\|1V}$ $[ps]$ | 87 | 87 | 119 (+37%) | 151 (+73%) |
| $D_{read\|0.35V}$ $[ns]$ | 6.36 | 6.36 | 7.69 (+21%) | 9.73 (+51%) |
| $E_{op\|1V}$ $[pJ]$ | 0.8 | 0.88 (+10%) | 0.86 (+7.5%) | 0.89 (+11%) |
| $E_{op\|0.35V}$ $[pJ]$ | 0.18 | 0.15 (-17%) | 0.23 (+30%) | 0.25 (+39%) |

We can notice that an automatic implementation may lead to an increase of 25% in area, more than 20% in delay and up to 50% in energy at ultra-low voltage. Custom implementation is thus highly pertinent for register file application.

### 4.5.3 Back biasing

Thanks to the FDSOI technology, we can apply a wide back biasing (symmetrical in this section) range on the register file transistors and observe the behaviour of the figures of merit.

For the same supply voltage, a forward body bias (FBB) decreases the threshold voltage, thus increases the speed and the leakage current. Figure 4.10 shows the energy per operation ($E_{op}$) effectively increasing for FBB at each supply voltage. As the leakage current is integrated over the entire clock period, the static energy at ultra-low voltage becomes higher than the dynamic energy and the energy-efficiency decreases dramatically. The supply voltage providing minimum $E_{op}$ varies between 0.3V and 0.5V, depending on the back bias.

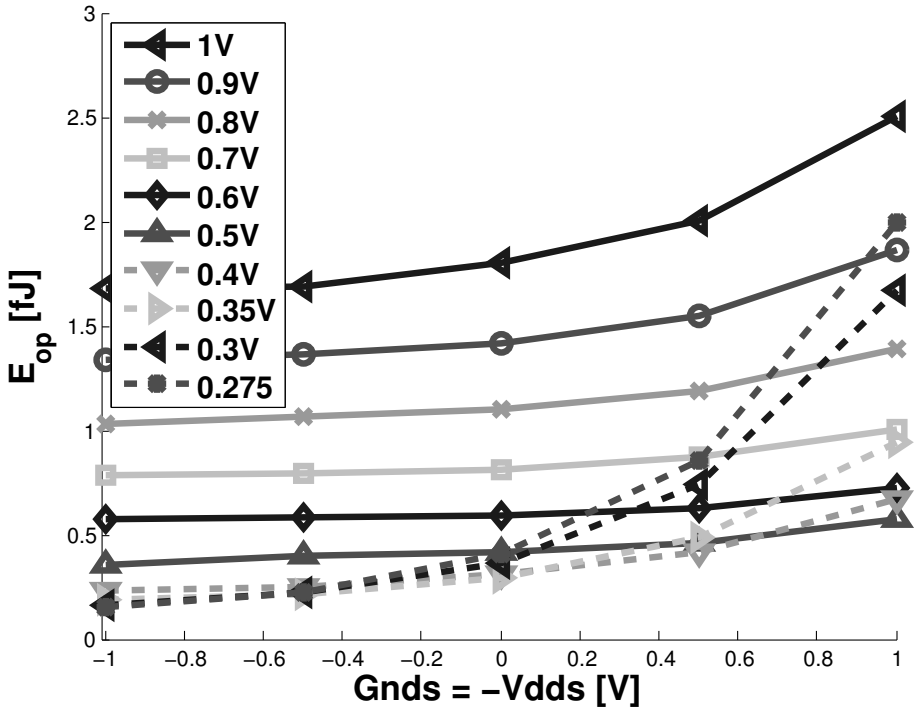



**Figure 4.10:** Energy per operation function of the supply voltage and the back biasing ($\alpha_{rate} = 15\%$, $T_{clk} = 40FO4$, corner TT, temp. = 25°C).

The figure 4.11 presents the evolution of the delay with the back bias. As with the silicon measurements in Chapter 2, the delay decreases with FBB and the relative variation is higher for low supply voltage. For strong FBB, the delay difference between two supply voltages reduces while this voltage reduction induces a lower dynamic energy consumption. For example, the delay at $V_{dd} = 0.7V$ is 73% higher than at $V_{dd} = 1V$ for Gnds = -Vdds = 1V, but the energy consumption is 2.5 times lower (figure 4.10).

This trend is even clearer in the figure 4.12, which represents the evolution of the energy-delay product (EDP), *i.e.* the combination of the two previous observations, with the supply voltage ($V_{dd}$) and back bias. For highest supply voltages, the energy consumption, impacted by the short-circuit current, mainly determines the shape. For the lower supply voltages, the delay is the most important part of variation. All combined, we can see that the minimum EDP is not reached at the same $V_{dd}$ for each back bias and that the lowest EDP is reached

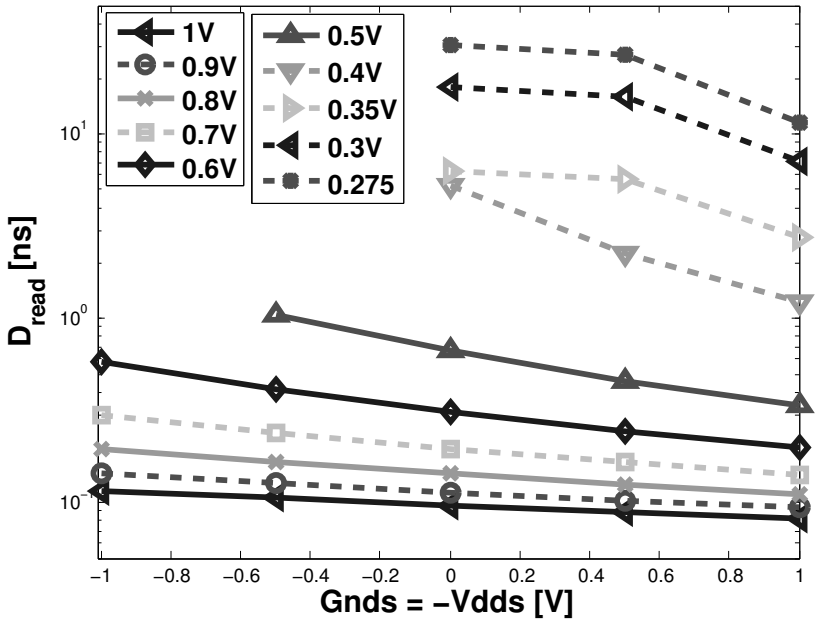



**Figure 4.11:** Evolution of the register file delay with back biasing.

for strong FBB. Therefore, we see that the energy-efficiency needs to combine adaptive dynamic supply voltage and back bias, as done in [24].

## 4.6 CONCLUSION

This chapter studied the integration of our energy-efficient pulse-triggered flip-flops at block level applications. First, we explained how our pulse-triggered flip-flops absorb the clock skew and allow to reduce this constraint during clock tree synthesis. After that, we presented a new conditional capture technique based on a pseudo-XOR gate which compares the data input and the current output. This pseudo-XOR gate is inserted at the beginning of the pulse generator which allows to save a lot of energy when data remains unchanged. A comparison in the energy-delay space has allowed the most energy-efficient variant to be selected depending on the targeted application. Then, we compared the architecture exhibiting the lowest $ED$ product with the pulsed-FFs of Chapter 2 and the conventional master-slave. We have shown that, thanks to our conditional capture technique, this new explicit pulse-triggered flip-flop architecture presents a lower delay but also a lower energy consumption than the master-slave architecture. As the problem of hold time is handled with the pseudo-XOR gate and

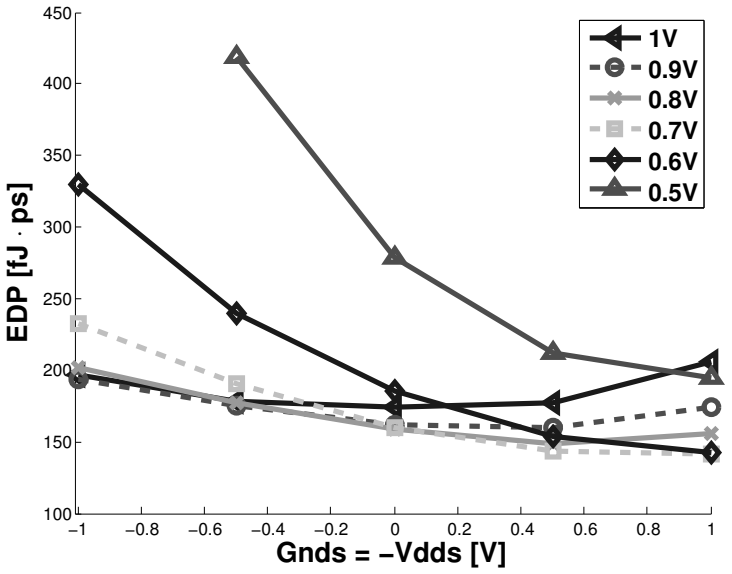



**Figure 4.12:** Evolution energy-delay product (EDP) with $V_{dd}$ and back bias.

the robustness is ensured by the current-starved technique (Chapter 3), we have designed an explicit pulsed-FF more energy-efficient than master-slave topology. Another advantage of this architecture is to provide another energy-delay trade-off for the CAD tools. Combined with the two pulsed-FFs architectures pointed out in Chapter 2 and the master-slave topology, the automatic synthesis tools are able to choose the lowest FF energy consumption while meeting the timing constraints of each path.

Afterwards, we studied the promising property of explicit pulsed-FFs: the pulse generator sharing. It has been shown that after a given number of latches sharing a PG, the energy consumption and area per FF are actually lower for explicit pulsed-FFs than for master-slave, even without pseudo-XOR gate. To highlight this idea, we implemented a structured register file with the energy-efficient explicit pulse-triggered flip-flop architecture developed in this work. It has been shown that the pulsed-FF-based register file effectively exhibits a lower energy per operation and lower area than the master-slave based register file.

# CONCLUSIONS AND PERSPECTIVES

In this dissertation, robust and energy-efficient explicit pulse-triggered flip-flop architectures targeting ultra-wide voltage range and ultra-low power circuits, have been developed and designed in FDSOI technology.

First of all, the explicit pulse-triggered flip-flop topology was pointed out of the literature in Chapter 1. This architecture presents interesting and remarkable timing properties, *e.g.* small input-to-output delay, negative setup time, time-borrowing technique, and dual-edge operation facilities, as well as shareable pulse generator. Nevertheless, this topology is hardly used in circuits working at ultra-low voltage because of two main drawbacks:

- the poor robustness to environmental variations compared to master-slave topology - handled in Chapter 3,
- the positive hold time which induces an energy overhead for inserting delay buffers - handled in Chapter 4.

In Chapter 2, an analysis of the latch operations led us to select six promising pulsed-FFs architectures for the targeted applications. Then, a fair comparison was performed in the energy-delay ($E-D$) domain to highlight the most energy-efficient architecture depending on the targeted application. If the TGPL-Clk architecture presents the best energy-efficiency for high-speed operations, the $C^2$MOS-Data architecture is revealed as the most energy-efficient pulsed-FF architecture over a wide range of targeted delays and supply voltage. Integration in a chip showed none additional difficulty to perform the timing closure. Then, silicon measurements showed the timing performances of flip-flops composed of the most energy-efficient latches and exhibited an average input-to-output delay down to 31ps at nominal voltage (1V), high temperature (80°C), and nominal back biasing conditions. Moreover, reverse and forward back bias allowed us to increase either the timing or the energy performances.

In Chapter 3, the fundamental tradeoff between robustness and energy consumption in a pulse-triggered flip-flop was explained and a current-starved delay generator (DG) was proposed to overcome this issue. It has been shown that our proposed current-starved DG significantly improves the robustness of the pulsed-FF, which was one of the two big drawbacks. Moreover, the structure is very flexible and offers many degrees of freedom to designers. Silicon measure-

ments compared the yield and minimum operating voltage of different flip-flops and showed that our proposed current-starved DG provides an average yield increase of 7.5% at the same $V_{dd}$ and an identical average yield reached at a supply voltage up to 45mV lower than without the current-starved DG. Moreover, it is explained how the proper choice of back bias can help designers to reach the lower operating voltage, or the higher yield. Then, an efficient approach of implementing the reset and enable functionalities in the pulse generator was presented.

In Chapter 4, it is shown that soft-edge pulse triggered flip-flops tolerate a large clock skew without modifying the delay performance. This absorption of the clock signal non-ideality facilitates the precision/energy tradeoff of clock tree synthesis. Afterwards, a conditional capture technique was presented and implemented in the most energy-efficient latch of Chapter 2. Then, a comparison in the $E-D$ domain showed that this pulse-triggered flip-flop architecture exhibits a lower energy dissipation than the master-slave topology and a negative hold time. The second drawback of pulsed-FF is thus partially alleviated thanks to this architecture. Finally, a 16x32bits register file, based on robust and energy-efficient pulsed-FFs coming from the development of the previous chapters and the pulse generator sharing property, is compared to MS-based register files. This comparison showed that our pulse-triggered flip-flop provides higher speed, lower area occupation and lower energy consumption, while guaranteeing a sufficient robustness in subthreshold regime.

In conclusion, we have designed energy-efficient pulsed-FF architectures, namely TGPL-Clk, C$^2$MOS-Data, and C$^2$MOS-Xor, respectively dedicated to high-speed, energy-efficient and ultra-low power operations. The FDSOI technology, through the back biasing technique, allows the energy-delay performances of these flip-flops to be dynamically modified, depending on the actual constraints of the circuit. The energy-efficiency is preserved at synthesis step thanks to the clock skew absorption, the conditional capture technique and during the addition of flip-flop functionalities thanks to the pulse generator structure. Finally, the robustness at ultra-low voltage is ensured by the low variability of FDSOI, the proper choice of the back bias value thanks to the yield study, and our proposed current-starved delay generator.

## Perspectives and future work

Three main pulsed-FFs architectures are pointed out from this work. The TGPL-Clk architecture is dedicated for high-speed operations, the C$^2$MOS-Data architecture is extremely energy-efficient over a wide range of supply voltage and timing performances, and the C$^2$MOS-Xor architecture presents a still lower energy consumption than MS and a negative hold time, which means no delay buffer insertion. By characterizing these architectures, we could synthesize a low-power microprocessor (the Cortex M0 for example) with our fast and energy-efficient flip-flops. Therefore, the expected energy saving comes from two ways. Firstly,

the clock period is reduced, so is the static energy per operation (the integration of the static current over the clock period) and finally the total energy per operation. Secondly, as the C$^2$MOS-Xor topology exhibits a lower energy consumption than master-slave flip-flops, it would provide more energy-efficient non-critical paths.

*Lógo mén* the C$^2$MOS-Xor will be used for the short paths, C$^2$MOS-Data in most of the cases and TGPL-Clk for the very critical paths ; *érgo dé* microprocessor architecture is so complex that the behaviour of synthesis is highly unpredictable, and the TGPL-Clk and C$^2$MOS-Data architectures could be used at the end of short paths. Thereby, their positive hold time would induce additional delay buffers for fixing the hold time constraint. Moreover, the clock load of the pulse-triggered structures is higher than the clock load of master-slave FFs of around 11% without shared PG. But, the relation between the total clock load and the clock tree consumption is not simple at all, and certainly not linear. For example, an industrial standard-cell library provides a finite amount of driving strength for the clock tree buffers. After a clock tree synthesis for a master-slave-based circuit, the sizes of the clock tree buffers are maybe already strong enough to drive the pulsed-FF clock load. Furthermore, the size of the buffers of the last branch of the tree might increase so much that they become as big as the previous branch, and an entire clock branch could be saved. Moreover, as already mentioned, the pulse generation sharing is an efficient way to reduce the energy consumption and the clock load by FF seen by the clock tree [81, 82, 83, 84]. Synthesis and Place&Route should be performed to study if the energy saving is higher or lower than the energy loss.

The structured datapath feature is a promising way for placing and routing a regular register file. While keeping the speed, energy and area performances of Chapter 4, it would add reconfigurable and reprogrammable properties to pulsed-FF-based register file.

As mentioned in Section 1.4, the time-borrowing technique is a very complicated microarchitectural system appeared in the literature these last few years. Here, it has the meaning of presenting a time-borrowing window after the triggering clock edge, sensing a valid data transition. If this transition is detected during the time-borrowing window, an error signal is generated and an error detection and/or correction mechanism(s) handle(s) this late data arrival. Yet, all the publications [52, 53, 54, 83] use master-slave and latch-based topology to perform the operations. As presented in Section 1.4, the implementation of the time-borrowing property needs less resources with pulse-triggered flip-flop than for master-slave-based topology.

Finally, the dual-edge property was discarded in Chapter 1 because we considered that the balance constraints on the clock tree seemed too difficult to reach at ultra-low voltage. A solution to overcome this problem might come from the interesting well properties of FDSOI technology. Thanks to the buried oxide (see Section 1.2.3), this technology allows PMOS and NMOS transistors to share an identical well. Indeed, the P-doped and N-doped channels can be over a buried oxide which is over a single P-well encircled by N-well. In [89], authors show

that this cell can be inserted into the logic with acceptable area penalty and leads to remarkable performances in fall/rise and propagation delays balancing and clock-tree skew. Moreover, an adaptive body biasing mechanism could be designed to sense and compensate the P-N imbalance and, contrary to [90], only one back bias generator would be needed for both PMOS and NMOS.

# REFERENCES


1. E. Beigne *et al.* "Ultra-Wide Voltage Range Designs in Fully-Depleted Silicon-On-Insulator FETs," Design Automation Test & Embedded (DATE), pp. 613-618, 2013.
2. D. Bol, "Pushing Ultra-Low-Power Digital Circuits into the Nanometer Era", Ph.D. Thesis, Louvain-la-Neuve, Belgique, December 2008.
3. G.E. Moore, "Cramming more components onto integrated circuits", Electronics, Vol. 38, No. 8, April 19, 1965.
4. S. Paik, G.-J. Nam, and Y. Shin, "Implementation of pulsed-latch and pulsed-register circuits to minimize clocking power," IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 640-646, Nov. 2011.
5. N. Magen *et al.* "Interconnect power dissipation in Microprocessor," International workshop on SLIP 2004.
6. M. Alioto, E. Consoli, G. Palumbo, "Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops : Part I-II", IEEE Transaction on Very Large Scale Integration (VLSI) Systems, vol. 19, no.5., pp. 737-750, May 2011.
7. J. Tschanz, S. Narendra, S. Borkar, M. Sachdev, and V. De, "Comparative Delay and Energy of Single Edge-Triggered and Dual Edge-Triggered Pulsed Flip-Flops for High-Performance Microprocessors", Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED), pp. 147-152. 2001.
8. V. Stojanovic, V.G. Oklobdzija "Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems," IEEE Journal of Solid-State Circuits, vol. 34, no. 4, 1999.
9. B. Rebaud, M. Belleville, C. Bernard, M. Robert, P. Maurine, and N. Azemard "A comparative study of variability impact on static flip-flop timing characteristics," IEEE International Conference on Integrated Circuit Design and Technology and Tutorial (ICICDT), pp. 167-170, Jun. 2008.
10. David Money Harris & Sarah L. Harris, "Digital Design and Computer Architecture", ed. Elsevier Inc, 2007.
11. H.J. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits", IEEE Journal of Solid-State Circuits, vol. 19 no. 4, pp.468-473, Aug. 1983.
12. H.J. Veendrick, "Nanometer CMOS ICs", ed. Springer, 2013.

13. M. Pelgrom, “Different faces of variability”, IEEE distinguished lecture, Universit catholique de Louvain (UCL), 2010.

14. K. Itoh & M. Horiguchi, “Low-voltage scaling limitations for nano-scale CMOS LSIs”, Solid-State Electronics, vol. 53, no. 4, pp. 402-410, 2009.

15. C.-Y. Chang *et al.*, “A 25-nm gate-length FinFET transistor module for 32nm node”, International electron device meeting (IEDM), 2009.

16. S.-Y. Wu *et al.*, “A 16nm FinFET CMOS technology for mobile SoC and computing applications”, International electron device meeting (IEDM), 2013.

17. L. Grenouillet *et al.*, “UTBB FDSOI transistors with dual STI for a multi-Vt strategy at 20nm node and below”, International electron device meeting (IEDM), 2012.

18. Q. Liu *et al.*, “High performance UTBB FDSOI devices featuring 20nm gate length for 14nm node and beyond ”, International electron device meeting (IEDM), 2013.

19. C. Fenouillet-Beranger, “Impact of a 10nm Ultra-Thin BOX (UTBOX) and Ground Plane on FDSOI devices for 32nm node and below”, Proceedings of European Solid-State Circuits Conference ESSCIRC, 2009.

20. D. Flandre, “Dispositifs électroniques avancés”, master course UCL, 2010.

21. S.A. Vitale *et al.*, “FDSOI Process Technology for Subthreshold-Operation Ultra low-Power Electronics”, Proceedings of the IEEE, vol. 98, No. 2, 2010.

22. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, “Sub-45nm Fully-Depleted SOI CMOS Subthreshold Logic for Ultra-Low-Power Applications”, SOI Conference, pp. 57-58, 2008.

23. V. Kilchytska, D. Flandre, F. Andrieu, “On The UTBB SOI MOSFET Performance Improvement In Quasi-Double-Gate Regime”, European Solid-State Device Research Conference (ESSDERC), 2012.

24. R. Wilson, E. Beigne, P. Flatresse, A. Valentian, F. Abouzeid, T. Benoist , C. Bernard, S. Bernard, O. Billoint, S. Clerc, B. Giraud, A. Grover, J. Le Coz, I.M. Panades, J.-P. Noel, B. Pelloux-Prayer, P. Roche, O. Thomas, Y. Thonnart, D. Turgis, F. Clermidy, P. Magarshack, “A 460MHz at 397mV, 2.6GHz at 1.3V, 32b VLIW DSP, embedding FMAX tracking”, the International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, pp. 452-453, 2014.

25. Weber O., Faynot O., Andrieu F., Buj-Dufournet C., Allain F., Scheiblin P., Deleonibus S. “High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding”, IEEE International Electron Devices Meeting, p.1-4, 2008

26. Jean-Philippe Noel *et al.*, “Multi-VT UTBB FDSOI Device Architecture for Low Power CMOS Circuit”

27. Y. Cheon P.-H. Ho A.B. Kahng S. Reda and Q. Wang “Power-aware placement,” international proceeding ACM/IEEE Design Automation Conference (DAC), pp. 795-800, 2005.

28. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, “An efficient metric of setup time for pulsed flip-flops based on output transition time”, Proceedings of the International Conference on Integrated Circuit Design and Technology (ICICDT), pp. 9-12, 2013.

29. U. Ko and P. T. Balsara "High-Performance Energy-Efficient D-Flip-Flop Circuits," Transactions on Very Large Scale Integration systems, vol. 8, no. 1, pp. 94-98, 2000.

30. D. Markovic B. Nikolic and R. Brodersen "Analysis and design of low-energy flipflops," in Proceedings of the International Symposium on Low Power Electronic Design (ISLPED), pp. 52-55, 2001.

31. H. Fuketa *et al.* "12.7-times Energy Efficiency Increase of 16-bit Integer Unit by Power Supply Voltage Scaling from 1.2V to 310mV Enabled by Contention-less Flip-Flops (CLFF) and Separated V DD between Flip-Flops and Combinational Logics", Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED), pp. 163-168, 2011.

32. C. Piguet "Low Power Electronics Design", CRC Press Computer engineering series, 2005.

33. D. Markovic, J. Tschanz, and V. De "Transmission-gate based flipflop," U.S. Patent 6 642 765, Nov. 4, 2003.

34. M. J. Myjak, J. G. Delgado-Frias, S. K. Jeon, "An Energy-Efficient Differential Flip-Flop for Deeply Pipelined Systems", 49th IEEE International Midwest Symposium on Circuits and Systems, pp. 203-207, 2006.

35. J. Yuan and C. Svensson "New single-clock CMOS latches and flipflops with improved speed and power savings," IEEE Journal of Solid-State Circuits, vol. 32, pp. 62-69, 1997.

36. B. Nikolic, V. Stojanovic, V.G. Oklobdzija, J. Wenyan, J. Chiu, and M. Leung, "Sense amplifier-based flip-flop," in Proceedings Solid-State Circuits Conference, pp. 282-283, Feb. 1999.

37. Y. Xia and A. Almaini "Differential CMOS edge-triggered flip-flop with clockgating", Electronics Letters, vol. 38, no. I, pp. 9-11, 2002.

38. C. K. Teh, M. Hamada, T. Fujita, H. Hara, N. Ikumi, and Y. Oowaki, "Conditional Data Mapping Flip-Flops for Low-Power and High-Performance Systems," Integration The VLSI Journal, vol. 14, no. 12, pp. 1379-1383, 2006.

39. W.-L. Su, H. Chiueh, P.-T. Huang, and W. Hwnag, "A Low Power Pulsed Edge-Triggered Latch for Survivor Memory Unit of Viterbi Decoder", 13th IEEE International Conference on Electronics Circuits and Systems, pp. 553-556, Dec. 2006.

40. M.-Y. Kim I. Jung Y.-H. Kwak and C. Kim "Differential pass transistor pulsed latch," Electrical Engineering, vol. 89, no. 5, pp. 371-375, 2006.

41. A. Ghadiri and H. Mahmoodi "Dual-edge triggered static pulsed flip-flops," 18th International Conference on VLSI Design held jointly with 4th International Conference on Embedded Systems Design, pp. 846-849, 2005.

42. B. Kong, S. Kim, and Y. Jun "Conditional-capture flip-flop for statistical power reduction," IEEE Journal of Solid-State Circuits, vol. 36, no. 8, pp. 1263-1271, 2001.

43. N. Nedovic "Conditional techniques for low power consumption flip-flops," Electronics Circuits, pp. 803-806, 2001.

44. P. Zhao *et al.*, "Low-Power Clock Branch Sharing Double-Edge Triggered Flip-Flop", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 15, no. 3, pp. 338-345, Mar. 2007.

45. W.M. Chung, "The Usage of Dual Edge Triggered Flip-flops in Low Power Low Voltage Applications", Ph.D. thesis, University of Waterloo, Ontario, Canada, January 2003.

46. P. Zhao T.K. Darwish and M.A. Bayoumi "High-Performance and Low-Power Conditional Discharge Flip-Flop," IEEE Transaction on VLSI Systems, vol. 12, no.5, pp. 477-484, May 2004.

47. M. Hamada *et al.* "A Conditional Clocking Flip-Flop for Low Power H.264/MPEG-4 Audio/Visual Codec LSI," Audio Visual, pp. 527-530, 2005.

48. C. K. Teh, T. Fujita, H. Hara, and M. Hamada "A 77% Energy-Saving 22-Transistor Single-Phase- Clocking D-Flip-Flop with Adaptive-Coupling Configuration in 40nm CMOS," the International Solid-State Circuits Conference (ISSCC), vol. 39, pp. 2010-2012, 2011.

49. E. Consoli, M Alioto, G. Palumbo, J. Rabaey, "Conditional Push-Pull Pulsed Latches With 726fJ.ps Energy-Delay Product in 65nm CMOS," the International Solid-State Circuits Conference (ISSCC) pp.482-484, 2012.

50. W.-H. Sung M.-C. Lee C.-C. Chung and C.-Y. Lee "Ultra-Low Voltage Implicit Multiplexed Differential Flip-Flop with Enhanced Noise Immunity", Electronics letters, vol. 48, no. 23, pp. 1452-1454, 2012.

51. T. Lin, C. Chien, and P. Chang, "A 0.48V 0.57nJ/pixel video-recording SoC in 65nm CMOS", International Solid-State Circuits Conference (ISSCC), pp. 158-160, 2013.

52. Ernst D. *et al.* "Razor : A Low-Power Pipeline Based on Circuit-Level Timing Speculation", Proceedings of the IEEE/ACM International Symposium on Microarchitecture, pp.7-18, 2003.

53. D. Blaauw, S. Kalaiselvan, K. Lai, S. Pant, C. Tokunaga, S. Das, D. Bull, "Razor II : In Situ Error Detection and Correction for PVT and SER Tolerance", the Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 400-402, 2008.

54. Chae K., Lee C., Mukhopadhyay S., "Timing error prevention using elastic clocking", International Conference on IC Design & Technology (ICICDT), pp. 1-4, 2011.

55. H.-M. Chou, H. Yu, and S.-C. Chang, "Useful-skew clock optimization for multi-power mode designs," IEEE/ACM International Conference on Computer Aided Design (ICCAD), pp. 647-650, Nov. 2011.

56. W. Shena and J. Hue, "Useful Clock Skew Optimization under A Multi-corner Multi-mode Design Framework", 11th International Symposium on Quality Electronic Design (ISQED), pp. 62-68, 2010.

57. D. Bol *et al.*. "SleepWalker: A 25-MHz 0.4-V Sub-mm$^2$ 7-$\mu$W/MHz Microcontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," IEEE Transaction Very Large Scale Integration (VLSI) Systems, vol. 48, no.1., pp. 20-32, Jan 2013.

58. V. Zyuban, "Optimization of scannable latches for low energy", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 11, no. 5, pp. 778-788, Oct. 2003.

59. D. Levacq V. Dessard and D. Flandre, "Low leakage SOI CMOS static memory cell with ultra-low power diode", in IEEE Journal of Solid-State Circuits, vol.42, no. 3, pp. 689-702, Mar. 2007.

60. T. Haine, F. Stas, D. Bol, "Optimization of the area / robustness / speed trade-off in a 28 nm FDSOI latch based on ULP diodes", Proceedings of Faible Tension Faible Consommation (FTFC), pp. 0-3, 2014.

61. J. P. Kulkarni, K. Kim, K. Roy, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM" IEEE Journal of Solid-State Circuits, vol. 42, no. 10, pp. 2303-2312, oct. 2007.

62. S. Bernard, "Etude et comparaison de cellules SRAM dual-port sans assist dynamique", master degree Université Catholique de Louvain, 2011.

63. I. Sutherland B. Sproull and D. Harris, "Logical Effort: Designing Fast CMOS Circuits", ed. Morgan Kaufmann, 1998.

64. M. Alioto, "Impact of NMOS/PMOS Imbalance in Ultra-Low Voltage CMOS Standard Cells", Proceedings 20th European Conference on Circuit Theory and Design (ECCTD), pp. 536-539, 2011.

65. J. Keane, H. Eom, T.-H. Kim, S. Sapatnekar, C. Kim, "Subthreshold Logical Effort: A Systematic Framework for Optimal Subthreshold Device Sizing", Proceedings of the 43rd annual conference on Design automation (DAC), pp. 425, 2006.

66. V. Zyuban and P. Strenski, "A Unified methodology for resolving power-performance tradeoffs at the microarchitectural and circuit levels", Proceedings of the 2002 international symposium on Low power electronics and design - ISLPED 2002, pp. 166, 2002.

67. M. Alioto E. Consoli G. Palumbo "General Strategies to Design Nanometer Flip-Flops in the Energy-Delay Space", IEEE Transaction on Very Large Scale Integration (VLSI) Systems, vol. 57, no.7, pp. 1583-1596, July 2010.

68. V. Stojanovic V.G. Oklobdzija R. Bajwa "A Unified Approach in the Analysis of Latches and Flip-Flops for Low-Power Systems", Proceedings of the international symposium on Low power electronics and design - ISLPED, 1998.

69. V.G. Oklobdzija, "Clocking and Clocked Storage Elements in a Multi-Gigahertz Environment", IBM J. Research and Development, vol.47, no.5/6, pp. 567–583, Sept. 2003.

70. S. Bernard, A. Valentian, D. Bol, J.D.-Legat, and M. Belleville, "A Robust and Energy Efficient Pulse Generator for Ultra-Wide Voltage Range Operations", Proceedings of the 5th Asia Symposium on Quality Electronic Design (ASQED), pp. 80-84, 2013.

71. B. Zhai S. Hanson D. Blaauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", Proceedings of the international symposium on low power electronics and design (ISLPED), pp. 20-25, August 2005.

72. X. Zhang and A. B. Apsel, "A low variation GHz ring oscillator with addition-based current source", Proceedings of the European Solid-State Circuit Research Conference (ESSCIRC), pp. 216-219, Sep. 2009.

73. P. Zhao *et al.* "Low-Power Clock Branch Sharing Double-Edge Triggered Flip-Flop", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 15, no. 3, pp. 338-345, Mar. 2007.

74. A. Ghadiri and H. Mahmoodi "Dual-Edge Triggered Static Pulsed Flip-Flops", 18th International Conference on VLSI Design held jointly with 4th International Conference on Embedded Systems Design pages, pp. 846-849, 2005.

75. D. Bol, D. Flandre, and J.-D. Legat, "Technology Flavor Selection and Adaptive Techniques for Timing-Constrained 45nm Subthreshold Circuits", Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design (ISLPED), pp. 19-21, 2009.

76. D. Bol, R. Ambroise, D. Flandre, J.-D. Legat, "Analysis and Minimization of Practical Energy in 45nm Suthreshold Logic Circuits", International Conference on Computer Design (ICCD), pp. 294-300, 2008.

77. M. Alioto, "Impact of NMOS/PMOS Imbalance in Ultra-Low Voltage CMOS Stdandard Cells", 20th European Conference on Circuit Theory and Design (ECCDT), pp. 536-539, 2011.

78. N. Nedovic, V.G. Oklobdzija, W.W. Walker, "A Clock Skew Absorbing Flip-Flop", ISSCC 2003.

79. D. Li, P. I.-J. Chuang, D. Nairn, M. Sachdev, "Design and Analysis of Metastable-Hardened Flip-Flops in Sub-Threshold Region", Proceedings of IEEE International Symposium Low Power Electronics and Design (ISLPED), Aug. 2011.

80. G.-P. Xiang, J.-Z. Shen, X.-X. Wu, and L. Geng, "Design of a low-power pulse-triggered flip-flop with conditional clock technique", Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 121-124, May 2013.

81. V. R. Ashna and M. Jagadeeswari, "Design of low power clocking system using merged flip-flop technique", Proceedings of the International Conference on Computer Communication and Informatics, pp.1-6, 2013.

82. H.-T. Lin, Y.-L. Chuang, Z.-H. Yang, and T.-Y. Ho, "Pulsed-Latch Utilization for Clock-Tree Power Optimization", IEEE Transactions on Very Large Scale Integration Systems, Issue 99, pp.1-13, 2013.

83. C. Chang and I. Jiang, "Pulsed-Latch Replacement Using Concurrent Time Borrowing and Clock Gating", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no. 2, pp.242-246,2013.

84. S. Paik, G.-J. Nam, Y. Shin, "Implementation of pulsed-latch and pulsed-register circuits to minimize clocking power", Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp.640-646, 7-10 Nov. 2011

85. P. Meinerzhagen and S. Sherazi, "Benchmarking of standard-cell based memories in the sub-VT domain in 65-nm CMOS technology", IEEE Transactions on Emerging and Selected Topics in Circuits and Systems, vol. 1, no. 2, pp.173-182, 2011.

86. F. Botman, J. De Vos, S. Bernard, J.D. Legat, D. Bol, "Bellevue: a 50MHz Variable-Width SIMD 32bit Microcontroller at 0.37V for Processing-Intensive Wireless Sensor Nodes", in IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1207-1210, 2014.

87. X. Zhao and Y. Ye, "Structure configuration of low power register file usingenergy model", Proceedings of IEEE Asia-Pacific Conference on ASIC, pp. 41-44, 2002.

88. M. Seok, D. Blaauw, and D. Sylvester, "Clock Network Design for Ultra-Low Power Applications", ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), pp. 271-276, 2010.

89. B. Giraud, J.P. Noel, F. Abouzeid, S. Clerc and Y. Thonnart, "Robust Clock Tree using Single-Well Cells for Multi-VT 28nm UTBB FD-SOI Digital Circuits", Proceedings of the SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), pp. 5-6, 2013.

90. Ragheb T., Ricketts A., Mondal M., Kirolos S., Links G. M., Narayanan V., and Massoud Y., "Design of thermally robust clock trees using dynamically adaptive clock buffers", IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 56, no. 2, pp. 374–383, Feb. 2009.

91. Lanuzza M., Rose R. De, Frustaci F., Perri S., and Corsonello P., "Comparative analysis of yield optimized pulsed flip-flops", Journal of Microelectronics Reliability, vol.52 no. 8, Aug. 2012 pp. 1679-1689.

# APPENDIX A

# ADDITIONAL STUDIES ON PULSE-TRIGGERED FLIP-FLOP ARCHITECTURES

The studies presented here do not give a fundamental message for the comprehension of the plain text. Nevertheless, some architectural choices in Chapter 2 may become clearer after this section.

First, we present the reasoning to reach the CDFF architecture compared in Chapter 2, which is slightly different from the original structure proposed in [46]. Second, simulations performed on the pulse-triggered version of the adaptive coupling flip-flop (ACFF [48]), the adaptive coupling pulse latch (ACPL), are shown to prove that it is not suitable for very-low voltage operations, as assessed in Section 2.2.2. Finally, in [6], the result of the comparison of pulsed-FFs showed that the CPFF architecture presents a lower energy consumption than TGPL in the low-power region of the energy-delay ($E - D$) domain. Here, we show that the resettable and scannable version of CPFF exhibits by far the worst energy-efficiency over all the pulsed-FFs.

### CDFF : keeper and reordering

From the original architecture proposed in [46], two points are analysed here: the keeper architecture and the order of the gate inputs in the two NMOS stack.

*Keeper* The aim of the feedback keeper of the original CDFF structure in [46], crossed in figure A.1, is to maintain the voltage value of the node when it has to stay high during the pulse signal. Keeper is needed since the leakage current and the period of the pulse signal are strongly affected by global and local variations

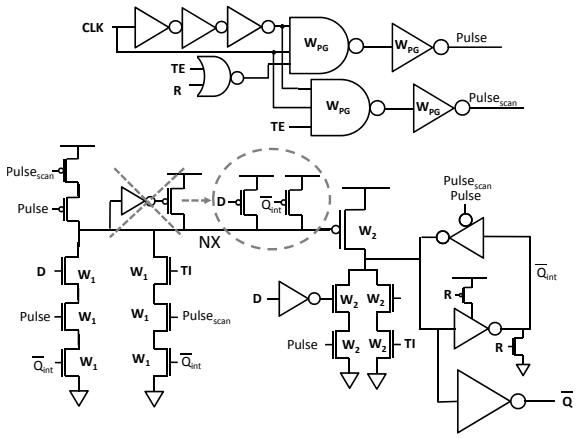



**Figure A.1:** Two-stages single-edge CDFF.

at ULV. Nevertheless, it creates a short circuit path when writing a logical 1 and two gates plus one drain junctions are added to NX.

As keeper should be active when either $D$ or $\overline{Q}$ or both equal zero, we propose a keeper composed of only two minimum-sized PMOS transistors connected to the intermediate node $NX$ (encircled in figure A.1). Thanks to that, no short circuit path is created during write time and the parasitic capacitance on $NX$ is reduced. The drawback is the increase of the data load but it is fully compensated by the removal of the short circuit path. It means that only a high number of glitches could get rid of the energy gain due to the absence of short circuit current.

Let us notice that, depending on the specifications and/or the technology, a keeper might be unnecessary. As a reminder, this structure needs a keeper mechanism to avoid the node NX to be discharged by leakage current when it should functionally stay at high level. This discharge causes additional energy overhead because NX will be in any case precharged after the triggering pulse signal. But, if the minimum level achieved, for the worst case of PVT and local variations, is sufficiently high, the keeper can be removed.Sufficient high means that the energy loss is lower than the energy gain due to the removal of the keeper and its parasitic capacitances, while improving the delay too.

*Reordering* Transistor reordering is a well-known technique that can be used to optimize circuit delay and power dissipation [91]. In the two NMOS stacks of the CDFF architecture (see figure A.1), the order of the input $D$ and $Pulse$ can be switched without changing the FF functionality.

The four configurations were previously compared and the results clearly showed that the configuration with the input data and its complementary $\overline{D}$ the closest to the stage output provides the best energy-efficiency for the whole energy-delay domain. That is because the intermediate junction capacitance has

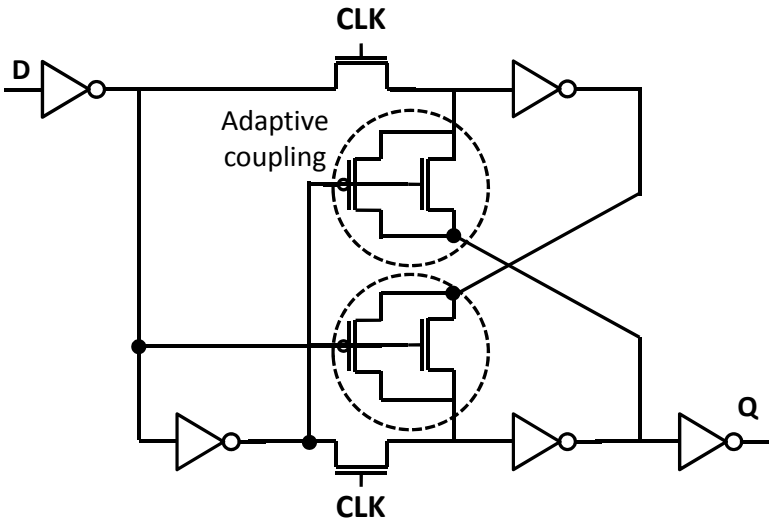



**Figure A.2:** As a reminder, the pulse-triggered version of the ACFF, the adaptive coupling pulse latch (ACPL). Even the basic architecture without additional functionalities is not functional at ultra-low-voltage.

been previously discharged and so the total amount of charge having to go to the ground, for the same transistor sizes, is reduced.

### ACPL

The ACPL topology (figure A.2) keeps the adaptive coupling technique proposed in [48], allowing a very small clock load. Nevertheless, this structure leads to floating nodes under certain conditions. Indeed, if the incoming data changes slightly after the end of the pulse signal, the inputs of the two cross-coupled inverters, providing the bistable element, are floating. Therefore, if the leakage current of the transistors gate-connected to the pulse signal is higher than the leakage current of the transistors gate-connected to the input data signal, the floating nodes might change their value, leading to a non correct functionality of the flip-flop.

Figure A.3 shows the results of 100 Monte-Carlo (MC) simulation made on the ACPL architecture. It shows that, at already 0.4V, the bistable element does not maintain the valid data value properly for 10% of the MC runs. On the other hand, if the gate width of the two pulsed transistors is not big enough, the writing system cannot switch the state of the bistable element during the pulse time. Therefore, the gate widths of both the pulsed transistors and the transistor performing the adaptive coupling technique (gate-connected to $D$) have to be

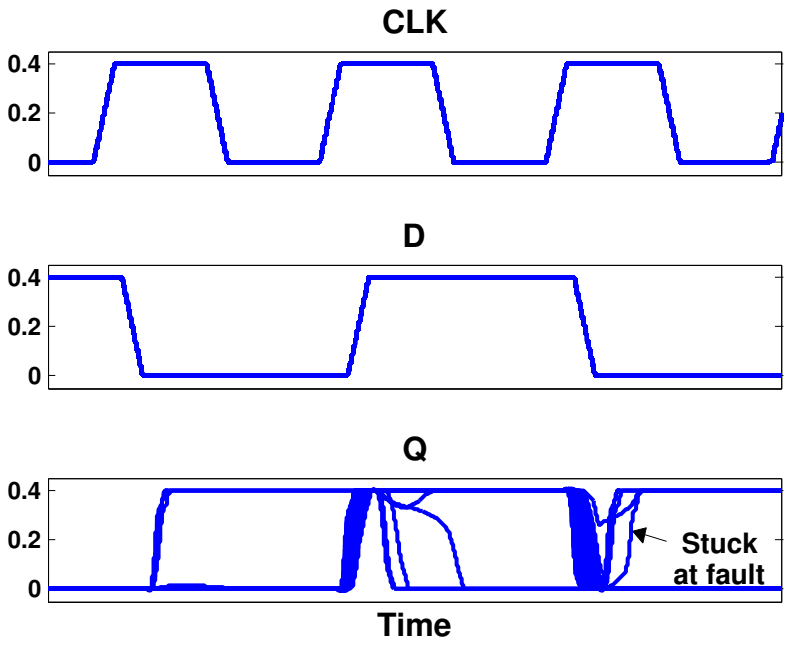



**Figure A.3:** Monte-Carlo simulations exhibit the weak robustness of ACPL architecture to transistor variability (corner FS, temp. = 25°C, $V_{dd}$ = 0.4V).

large enough to guarantee sufficient writing and leakage currents, respectively. As large gate widths, meaning high energy consumption, are needed for ensuring a correct functionality, this architecture has not been selected for the comparison.

### Resettable and scannable CPFF

From the large study and comparison of [6], the conditional precharge flip-flop (CPFF) architecture presents the best energy-efficiency in low power domain for small data activity among all studied pulse-triggered flip-flops. However, all the studied structures in [6] only perform the basic flip-flop functionality, *i.e.* with only three (D, CLK, Q) ports. In this section, the reset and scan property is added to the CPFF architecture (figure A.4), as the other pulsed-FFs.

Figure A.5 shows the comparison in the $E - D$ space of the pulsed-FFs architectures. We see that the resettable and scannable CPFF architecture is overwhelmingly the worst energy-efficient without presenting good timing performances. Because of its implicit pulse property, as well as its feedback system, a quite large amount of additional inverters is needed to perform both reset and scan functionalities. Therefore, CPFF is not more energy-efficient than TGPL topology in low-power region, in contradiction with the results in [6]. As already said, this shows that adding FF functionalities may change the comparison result and should be included in every comparison of FF in advanced technology.

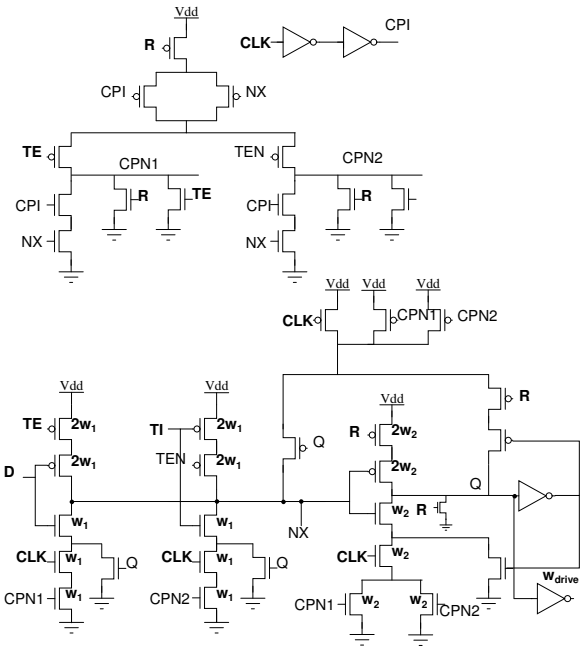



**Figure A.4:** R-S CPFF architecture.

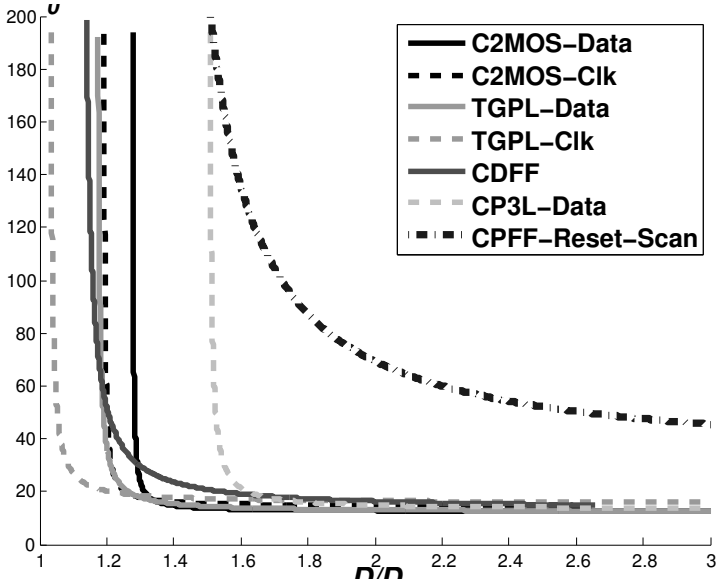



**Figure A.5:** EECs with CPFF architecture (driving strength X1, $V_{dd} = 1.0$V; $\alpha_{sw} = 0.15$, $T_{clk}/FO4 = 40$, corner TT, temperature 70°). Additional transistors lead to a very poor $E - D$ performances.

# APPENDIX B

# TESTBENCH FOR FLIP-FLOP AND ENERGY-DELAY ESTIMATION

In this section, we present in detail the testbench used for every flip-flop comparison in this work, the definition of the delay and energy, and the sizing methodology which allowed us to get/compute the energy-efficient curves (EECs) described in Section 2.3.

### Testbench

Our testbench, represented in figure B.1, is largely inspired from [6]. The data and clock input signals have a FO3 inverter slope which is tuned for every environmental conditions. The reset and scan signals are held in disable mode, *i.e.* $TE = 0V$ and $RN = V_{dd}$ or $R = 0V$. The scan input data is the inverse value of the input data $D$ in order to take the worst case, thus maximum delay, into account. The output load is composed of three identical inverters of same driving strength as the FF under test and each of those three inverters is connected to another inverter to avoid unrealistic Miller effect. This emulates a FO3 output, quite common after synthesis, and takes into account the dependency of the transistor gate capacitance with the gate voltage. The current going from the FF to the load is integrated and then removed from the total energy computation. The energy result is thus load independent.

Since [6] showed that they largely influence the FF architectures comparison, layout parasitics are evaluated and taken into account in the design methodology. Their evaluation is performed thanks to the geometrical approach proposed in [67].

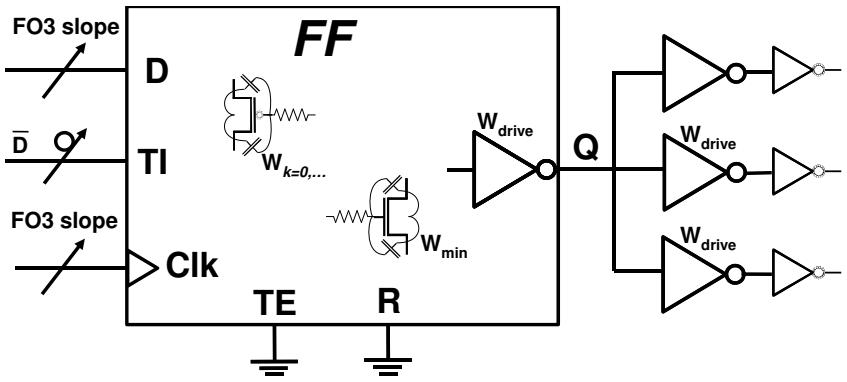



**Figure B.1:** Testbench used to characterize FFs with layout parasitics included.

### Definition of timing parameter and energy consumption

In this dissertation, the delay of the flip-flops are defined by the data-to-output (D-to-Q) delay computed at setup time. The metric used to compute the setup time is a new metric we proposed, based on the output signal transition time [28]. Compared to the state of the art, it really gives the limit beyond which the performance of the flip-flop is degraded and the reliability is endangered for pulsed-FFs. As it takes into account the soft-edge property of pulsed-FFs, this metric provides the most timing efficient value for the setup time of pulsed-FF and allows the maximal clock frequency obtained by synthesis to be reduced. Here, the setup time is by definition the time between the data edge and the triggered clock edge such as the transition time of the output increases of 10% compared to the transition time of the output when input data arrives far before the clock edge.

The average FF energy consumption by clock cycle depends on many parameters, including the data input switching activity ($\alpha_{sw}$), the clock period ($T_{clk}$) and the temperature. As proposed in [6] in order to be technologically independent, $T_{clk}$ is normalized by the FO4 inverter delay, giving the logic depth of the circuit, and delay and energy are normalized respectively by:

- $D_0$, the $FO4$ inverter chain delay in the same environmental conditions,
- $E_0$, the energy dissipated by an unloaded symmetrical minimum sized inverter during a complete $0 \rightarrow 1 \rightarrow 0$ transition cycle.

The total energy dissipation is computed as described in the appendices of [6], where the output current is removed from the total FF consumption and the effect of every input transitions is taken into account, as well as input data and clock loads.

### Design methodology

Our sizing methodology is also largely inspired from [6]. Only the gate widths of the transistors in the D-to-Q path ($W_k$) can modify the speed, and thus the $E - D$ tradeoff. Thereby, they are the main variables for the transistor sizing algorithm. This algorithm consists in determining the set of $W_k$ which provides a minimum point for a given figure of merit (FOM). Since modern applications of digital electronic range from high-speed to low-power designs, a large class of FOMs $E^i D^j$ have been adopted to cover all the possible tradeoffs. From a discrete set of design points minimizing several $E^i D^j$ FOMs, the energy-efficient curve (EEC) of the FF, *i.e.* the set of design points showing minimum energy (delay) for a given delay (energy) [66], can be extracted. We chose to consider $ED^j$ and $E^i D$, for $i, j = 1...5$, because they cover a very wide range of applications. To reduce the total number of variables in the sizing algorithm, some simplifications are introduced:

- series-connected transistors are equally sized for litho-friendly layout [12],
- pull-up and pull-down network in the D-to-Q path are symmetrically sized (see Section 2.3.1),
- transistors of the pulse generator ($W_{PG}$) are sized in order to achieve a FO3 fall and rise time for the pulse signal,
- on the contrary to [6], the gate width of the output inverter ($W_{drive}$) is an input of the sizing algorithm in order to be more realistic compared to an industrial standard-cell library,
- the rest of the transistors are minimally sized.

# APPENDIX C

# TESTBENCH FOR REGISTER FILE

The testbench used to compute the delay and energy performances of the register files was developed with a commercial electrical simulator, using the RCc extractions of the register file layouts.

As a reminder, the input-output ports for each $16 \times 32b$ register file are:

- [1 bit] clock, write-enable, reset, scan-enable,
- [4 bits] read-addr1, read-addr2, write-addr,
- [32 bits] write-data, write-data-scan,
- [output 32 bits] read-data1, read-data2

Similarly to the testbench in Annex B, each of the input presents a FO3 input slope and each of the output is connected to three identical inverters of the same drive as the FFs inside the register file (see figure C.1).

The clock signal is generated thanks to a FO3 input slope on the gate of the first inverter (or buffer) of the clock tree. Notice that the input clock signal is not connected directly to the FFs. The write-enable signal rises to $V_{dd}$ during a writing cycle and then goes back to *gnd*. It allows us to take into account the dynamic energy induced by the write-enable signal during writing operations. After a global reset at the beginning of the simulation, the reset signal is tied to inactive value, as well as the scan-enable signal.

The write-addr signal changes from 0110 to 1001 during the writing cycle, thus performing a maximum activity, and is then maintained during the rest of the operations. During the first reading cycle, read-addr1 passes from 0110 to 1001 and the read-data1 output is sensed. In addition to performing a maximum data activity for the energy computation, we compute the read-access-time1 which is by definition the worst propagation time between the read-address change and the output signal switching. In the second reading cycle, read-addr2 goes from 0110 to 1001 and both the read energy and access-time is computed.

The write-data signal is a random chain of 32bits with equal number of 0 and 1. None of the 32 input bits changes in the first writing cycle and all of them vary

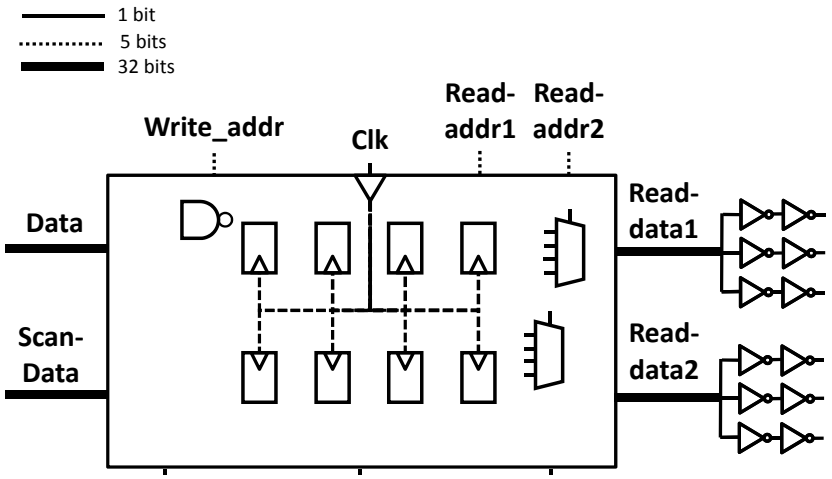



**Figure C.1:** Testbench used to characterize the register files (post-layout extraction).

in the second cycle. This emulates a 0% and 100% data input rate activity factor ($\alpha_{rate}$), the probability of the inputs to flip their state during a clock cycle. The 32 bits of write-data-scan signal are kept at constant values.

Finally, the delay is defined as the worst read-access-time between read-access-time1 and read-access-time2.

$$D_{read} = max(D_{read,A,i}, D_{read,B,i}) \tag{C.1}$$

where $i = 0, ..., 31$. The writing-access is not computed because it would have needed a bisection method with several iterations to defined the setup time. Because of the huge netlist provided after layout extraction, the needed computation time would have been too high. Nevertheless, the inputs are directly connected to the FF inputs and we know from previous chapters that the D-to-Q performance of pulsed-FFs is highly superior of the timing performance of master-slave flip-flops.

The energy per operation is the combination of several computed energies coming from several clock cycles. In the first clock cycle, a global reset is performed followed by a writing operation with known data values. No energy is computed. In the second clock cycle, the write-enable signal is activated but a zero data activity is seen on the inputs. The energy is the integration of the supply plus input currents: $E_{w0}$. In the third clock cycle, a writing with 100% data rate activity is performed, giving $E_{w100}$. In the fourth clock cycle, we read a new set of 32 bits from the output signal 1 $E_{r1}$. Because of the 100% activity imposed in the previous cycle, all the read output values switch after the read-address

change. And in the last clock cycle, we read a complete new set of 32 bits from the output signal 2 $E_{r2}$ where, again, every output switches.

As a read operation is asynchronous, we only computed the read-energy until the end of the output switch. Indeed, the supply current $I_{dd}$ is not integrated over the whole read clock cycle since the leakage energy has already been taken into account in the write-energy computation. More precisely, the end of the integration is the time when:

$$\frac{dI_{dd}}{dt} < 1000 \quad \& \quad (I_{dd} < 0.1 I_{dd,max} \quad || \quad I_{dd} < 2 I_{dd,end})$$

where $I_{dd,max}$ is the maximum supply current reached during the read clock cycle and $I_{dd,end}$ the supply current computed at the end of the read clock cycle.

Finally, the energy per operation is defined as:

$$E_{op} = \alpha_{rate} E_{w100} + (1 - \alpha_{rate}) E_{w0} + \alpha_{rate} (\frac{E_{r1} + E_{r2}}{2}) \tag{C.2}$$

# APPENDIX D

# RÉSUMÉ EN FRANÇAIS

Avec l'explosion du marché des applications portables et le paradigme de l'Internet des objets, la demande pour les circuits à très haute efficacité énergétique ne cesse de croître. Afin de repousser les limites de la loi de Moore, une nouvelle technologie est apparue très récemment dans les procédés industriels afin de remplacer la technologie en substrat massif ; elle est nommée *fully-depleted silicon on insulator* ou FDSOI.

Dans les circuits numériques synchrones modernes, une grande portion de la consommation totale du circuit provient de l'arbre d'horloge, et en particulier son extrémité : les bascules. Dès lors, l'architecture adéquate de bascules est un choix crucial pour atteindre les contraintes de vitesse et d'énergie des applications basse-consommation. Après un large aperu de l'état de l'art, les bascules à impulsion explicite sont reconnues les plus prometteuses pour les systèmes demandant une haute performance et une basse consommation. Cependant, cette architecture est pour l'instant fortement utilisée dans les circuits à haute performance et pratiquement absente des circuits à basse tension d'alimentation, principalement à cause de sa faible robustesse face aux variations.

Dans ce travail, la conception d'architecture de bascule à impulsion explicite est étudiée dans le but d'améliorer la robustesse et l'efficacité énergétique. Un large panel d'architectures de bascule, avec les fonctions *reset* et *scan*, a été comparé dans le domaine énergie-délais, à haute et basse tension d'alimentation, grce à une méthodologie de dimensionnement des transistors. Il a été montré que la technique dite de polarisation face arrière, l'un des principaux avantages de la technologie FDSOI, permettait des meilleures performances en énergie et délais que la méthodologie de dimensionnement. Ensuite, comme le générateur d'impulsion est la principale raison de dysfonctionnement, nous avons proposé une nouvelle architecture qui permet un très bon compromis entre robustesse à faible tension et consommation énergétique. Une topologie de bascule à impulsion explicite a été choisie pour être implémentée dans un banc de registres et, comparé aux bascules maître-esclave, elle présente une plus grande vitesse, une plus faible consommation énergétique et une plus petite surface.

## D.1 INTRODUCTION

Les applications portables, telles que les *smartphones*, les tablettes et les téléphones portables, ainsi que les circuits à très basse consommation, tels que les puces RFID, les réseaux de capteurs sans fils et les applications biomédicales, portent vritablement l'industrie de la micro électronique aujourd'hui. Dans ces applications, le microprocesseur est connecté à une batterie ou un système de récupération d'énergie, ce qui signifie, dans les deux cas, une énergie limitée d'énergie et de puissance disponible. Par conséquent, l'efficacité énergétique est d'une importance capitale pour la conception de ce type de circuits.

La technologie appelée FDSOI est apparue récemment dans les procédés industriels, afin de surmonter les limites de la technologie en substrat massif. Grâce à son meilleur contrôle électrostatique du canal des porteurs, cette technologie apporte une plus faible capacité de jonction, une pente sous-seuil plus raide, une plus faible variabilité et une très performante technique : la polarisation face arrière sur une très large gamme de tensions. Cette polarisation face arrière permet de modifier dynamiquement la tension de seuil des transistors de manière réversible. En conséquent, cette technologie convient parfaitement aux circuits à haute efficacité énergétique et très basse consommation.

Dans les circuits numériques synchrones modernes, le nombre de bascule a littéralement explosé avec la montée en puissance de nouvelles techniques micro-architecturales. Ainsi, l'architecture de bascule a un rôle et un impact décisif sur les performances temporelles ainsi que la consommation énergétique du processeur. Les bascules à impulsion explicite présentent de remarquables caractéristiques temporelles, permettant de gagner une part non-négligeable du cycle d'horloge. En même temps, sa consommation énergétique peut être drastiquement réduite par le partage du générateur d'impulsion. Jusqu'alors, cette structure est presque complètement absente des circuits travaillant à très basse tension d'alimentation, pour lesquelles les architectures de type maitre-esclave sont principalement utilisées. Les deux principales raisons en sont :

- Une plus faible robustesse face aux variation local dans la génération de l'impulsion,
- Un temps de maintien (*hold time* en anglais) positif, provoquant des tampons en délais additionnels et donc une surconsommation énergétique.

Dans le but d'améliorer les performances des circuits à haute efficacité énergétique, d'un point de vue à la fois en vitesse et en consommation, ce travail étudie et analyse des innovations architecturales pour surmonter et résoudre ces deux désavantages.

L'étude est menée par les deux questions suivantes :

- Comment obtenir des bascules robustes et efficaces en énergie à très basse tension d'alimentation ?
- Comment la technologie FDSOI peut-elle nous aider à améliorer la robustesse et l'efficacité énergétique ?

Le manuscrit est articulé autour de quatre chapitres principaux qui sont repris de manière succincte dans ce résumé étendu :

Dans le chapitre 1, la technologie FDSOI et tous ses avantages sont présentés en détails. Ensuite, nous dressons un état de l'art des quatre topologie de bascule CMOS. Les configurations maitre-esclave, différentielle, à impulsion et à double-front sont illustrées avec des architectures de bascule de la littérature scientifique. Pour chacune d'entre elles, les avantages et désavantages sont expliqués et ensuite résumés. A la fin du chapitre, nous pointons la structure de bascule à impulsion explicite comme candidate prometteuse pour augmenter les performances des circuits à haute efficacité énergétique. Cette topologie souffre néanmoins de deux principaux inconvénients, à savoir la robustesse et le temps de maintien positif, qui sont traités dans le chapitre 3 et 4, respectivement.

Dans le chapitre 2 sont comparées six architectures prometteuses pour nos applications, dans le plan énergie-délais. L'architecture appelée $C^2$MOS-Data se révèle comme la plus efficace énergétiquement sur une large gamme de tension d'alimentation. Après cela, nous montrons comme la polarisation face arrière, permise par la technologie FDSOI, peut apporter de meilleures performances en énergie et délais que la méthodologie de dimensionnement utilisées précédemment. Enfin, des mesures sur silicium viennent confirmer les résultats précédemment obtenus.

Le chapitre 3 commence par expliquer le compromis inhérent des structures à impulsion à très basse tension d'alimentation. Pour assurer une robustesse suffisante, une surconsommation énergétique est payée sous plusieurs formes. Pour contourner ce problème, nous proposons une nouvelle architecture de générateur de délais (GD), basée sur la technique du *current-starved*, présentant assez de degrés de liberté pour atteindre la robustesse désirée sans pénalité de consommation. Ensuite, des simulations *post-layout* et des mesures sur silicium montrent comment notre structure de GD améliore significativement la robustesse des bascules à impulsion.

Dans le chapitre 4 est premièrement présentée et expliquée une technique de capture conditionnelle. Sachant le générateur d'impulsion (GI) est le plus grand consommateur d'énergie au sein des bascules à impulsion, cette technique attaque très efficacement la consommation énergétique de la bascule. Par ailleurs, les courbes d'efficacité énergétique montrent que l'énergie par opération obtenue grâce à cette technique est plus petite que celle des architectures maitre-esclaves. Ensuite, après avoir confirmé l'efficacité du partage de la génération d'impulsion, nous intégrons des innovations précédemment exposées, à savoir le rapide et énergie-efficace *latch*, le robuste et énergie-efficace générateur de délais et le GI partagé, afin d'implémenter un banc de registre. Ce banc de registre basé sur des structures de bascules à impulsion présente une plus grande vitesse, une plus faible consommation énergétique et une plus petite surface qu'un banc de registre basé sur des bascules maitre-esclave.

Finalement, nous terminons par une conclusion résumant ce travail.

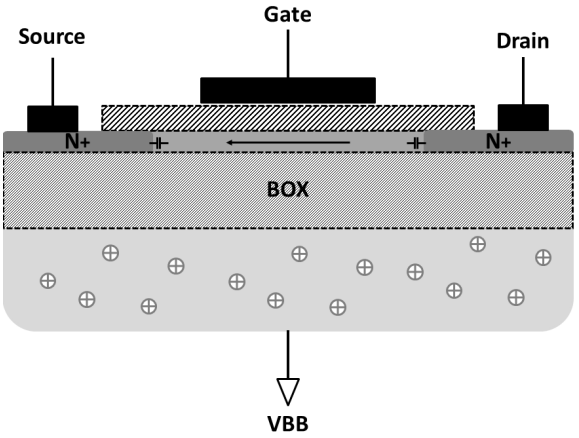


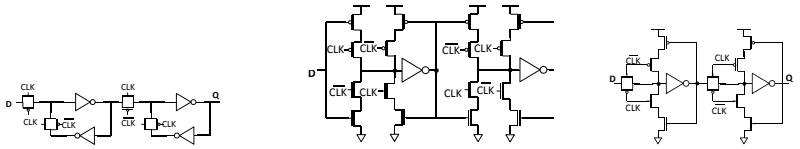

**Figure D.1:** Le transistor FDSOI (version NMOS).

### D.1.1 La technologie FDSOI

Le transistor FDSOI est représenté à la figure D.1. Une couche mince d'oxyde, ou oxyde enterré - appelé BOX sur la figure D.1 -, est insérée entre le substrat et la partie active du transistor, le canal des porteurs. Ainsi, la hauteur du canal n'est que de quelques nanomètres (Pour le nœud 28nm, 8nm de silicium sont déposés au-dessus de 25nm d'oxyde enterré). Contrairement à la technologie en substrat massif, la région en dessous de l'oxyde, la face arrière du transistor, n'est plus nécessairement maintenue à une tension d'alimentation. Quand la tension d'activation du transistor (tension grille-source pour le NMOS) est plus grande que la tension de seuil, le canal des porteurs est, grâce à sa très mince hauteur, complètement déplété. Cette propriété a donné son nom à la technologie, *fully-depleted silicon on insulator* (FDSOI) en anglais, ou en français, silicium sur isolant totalement déplété.

Cette configuration apporte plusieurs avantages :

- Une capacité de jonction fortement réduite [22, 23].
- Un meilleur contrôle électrostatique de la grille et de la face arrière sur le canal. Cela a pour conséquence une meilleure pente sous-seuil et un meilleur effet de substrat, ainsi que la réduction des effets canaux courts, comme le DIBL.
- Une disparition de la jonction PN entre les sources et drain et le substrat. Cela signifie que nous ne sommes plus limité aux 0,3V de différence de tension comme en technologie en substrat massif. Il a été montré sur silicium qu'il est possible d'appliquer une tension de substrat jusqu'à 2 volts sans claquage électrique [24]. De plus, cela signifie que l'on peut, durant la vie du circuit, dynamiquement varier cette tension de substrat et donc la tension de seuil et les propriétés électriques des transistors.




**(a)** L'architecture maitre-esclave d'origine à porte de transmission.

**(b)** L'architecture complémentaire CMOS ($C^2MOS$). Le phénomène de *pass-through* est évité.

**(c)** L'architecture conventionnelle de bascule à porte de transmission.

**Figure D.2:** Architectures maitre-esclave conventionnelles. TGFF donne un meilleur compromis entre la vitesse et le phénomène *pass-through*.

- Contrairement à la technologie en substrat massif, il n'y a plus d'atome dopant dans le canal du transistor FDSOI. Grâce à cela, la variabilité intrinsèque du transistor est fortement réduite [25]. De plus dans [26], les auteurs expliquent comment obtenir trois tensions de seuil différentes (pour NMOS et PMOS) en jouant sur le dopage de la face arrière. Il est donc possible d'obtenir la même diversité qu'en substrat massif, tout en conservant une faible variabilité.

### D.1.2 État de l'art des bascules

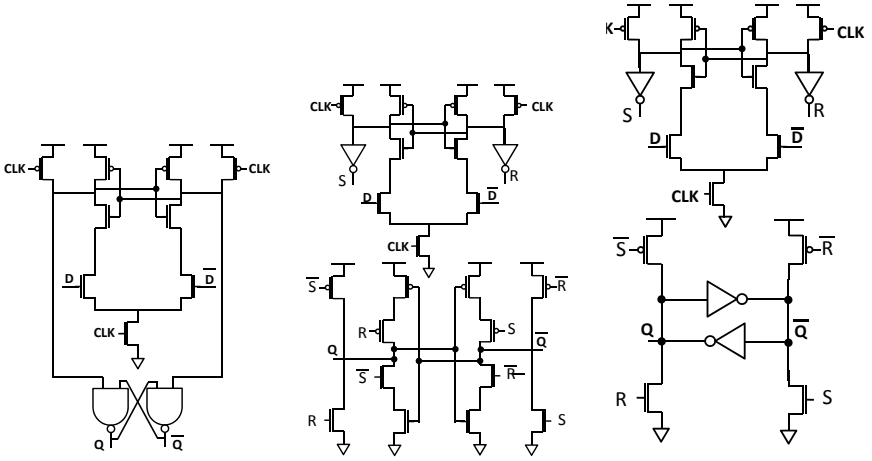
La bascule est l'élément fondamental des circuits numériques synchrones. Ces dernières années, leur nombre au sein des circuits a littéralement explosé du fait de nouvelles techniques microarchitecturales, comme le *pipelining*, les architectures super-scalaires et les techniques de *time-borrowing*. De ce fait, les bascules ont un rôle déterminant sur la vitesse, l'énergie, la surface et la robustesse des circuits. On estime que 50% à 70% de la consommation totale provient de l'arbre d'horloge [4, 5], et 80% de l'énergie dynamique de l'arbre est située au niveau des bascules [27].

Ainsi donc, l'architecture des bascules est un élément primordial que les concepteurs de circuits doivent prendre en compte pour concevoir un circuit à haute efficacité énergétique.

Nous allons maintenant présenté l'état de l'art des architectures de bascule présentes dans la littérature scientifique. Les bascules peuvent être cataloguées en quatre grandes catégories : maitre-esclave, différentielle, à impulsion, double-front.

#### *D.1.2.1 Maitre-esclave*

Les bascules maitre-esclave sont composées de deux *latch*, non nécessairement identiques, placés en série (voir figure D.2. La donnée d'entrée (D) est connecté au *latch* maitre qui est actif sur niveau bas de l'horloge. Lors du front montant,




**(a)** Structure basique de bascule SAFF [35]. **(b)** Topologie améliorée de SAFF [36]. **(c)** Avec l'étage de sortie modifié.

**Figure D.3:** Bascules différentielles basées sur le *sense-amplifier*.

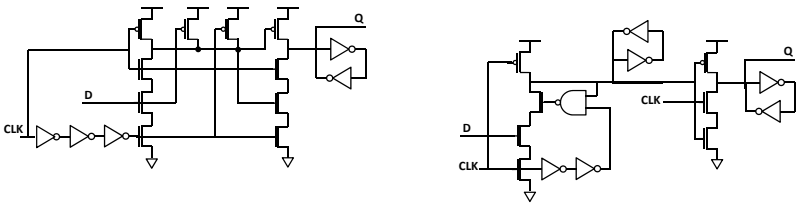
le *latch* maitre devient bloqué et transmet l'information de la donnée au *latch* esclave devenant passant. Ce dernier transmet l'information à la sortie (Q) de la bascule pour l'étage suivant du circuit et la maintiendra jusqu'au prochain front montant de l'horloge.

Le temps de *setup* dépend donc principalement de la vitesse de commutation du maitre, tandis que le temps de propagation entre les fronts d'horloge et de sortie (Clk-Q) dépend de l'esclave. Le temps de maintien est, pour des conditions d'utilisation normales, toujours positif car comme le maitre est fermé après le front de l'horloge, une variation de l'entrée n'influe en rien la sortie.

Le temps de propagation entre l'entrée et la sortie (D-Q) est la somme du temps de *setup* plus le Clk-Q. La donnée devant passer deux *latch*, ce type de bascule est relativement lent par rapport aux autres catégories. De plus, le temps D-Q explose lorsque l'on diminue la tension d'alimentation. Les bascules de type maitre-esclave ne semblent donc pas appropriées pour la haute performance à très faible tension.

### *D.1.2.2 Différentiels*

Les architectures typiques de bascules différentielles sont représentées à la figure D.3. Ces bascules présentent à la fois la sortie logique ($Q$) et son complémentaire ($\overline{Q}$) en sortie et nous pouvons remarquer une très forte symétrie axial dans la structure.




**(a)** Bascule à *latch* hybride (HLFF) [8]. **(b)** Bascule semi-dynamique (SDFF) [8].

**Figure D.4:** Bascule semi-dynamiques à impulsion implicite.

Ces bascules sont extrêmement rapides, surtout comparées aux maitre-esclave. Néanmoins, elles nécessitent une pré-charge à chaque cycle d'horloge, quelque soit la valeur de la donnée. Cela augmente sensiblement la consommation énergétique et n'est pas adapté à la haute efficacité énergétique. De plus, elles sont très peu robustes lorsque l'on diminue la tension d'alimentation. En effet, il y a toujours un nœud auquel une pile de transistor NMOS est connectée et, lors de l'écriture, tente de décharger ce nœud pendant que un ou plusieurs transistors PMOS tentent de le maintenir à la tension d'alimentation. La réduction de la tension d'alimentation, combinée aux variations de procédé de fabrication, rend cette catégorie de bascule caduque pour notre application.

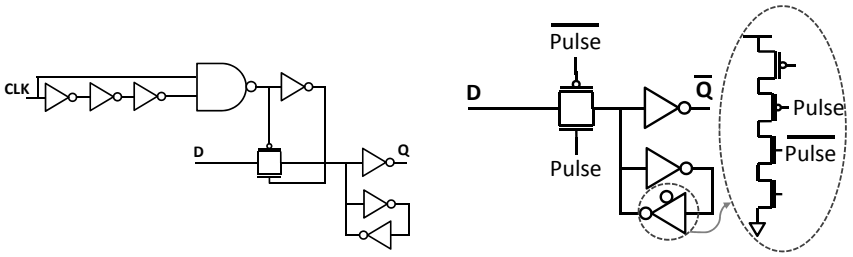
#### *D.1.2.3 A impulsion*

Les bascules à impulsion sont composées d'un seul *latch* ouvert pendant une courte période après le front montant de l'horloge. Elles ne fonctionnent pas à proprement parlé sur un front d'horloge, mais pendant une impulsion suivant le front. Ce comportement est obtenu grâce à l'utilisation d'un signal d'horloge retardé, dont le retard définit la largeur de l'impulsion. De part l'architecture, si la donnée change légèrement après le front d'horloge, elle peut être échantillonnée par la bascule. Cela signifie que le temps de *setup* est négatif, mais en contrepartie, le temps de maintien est positif.

Au sein de cette catégorie, il existe deux grandes familles : les implicites et les explicites.

Les bascules à impulsion implicite ont leur génération d'impulsion intégrée directement dans le *latch* (voir figure D.4). Cette génération intégrée permet de consommer légèrement moins que leurs pendants explicites, mais néanmoins, les architectures implicites nécessitent un minimum de trois transistors empilés pour transmettre l'information de l'entrée à la sortie.

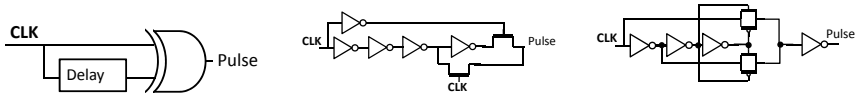
Les architectures explicites ont un générateur d'impulsion qui peut être facilement discerné de l'architecture du *latch* (voir figure D.5). Recevant le signal d'horloge, le générateur envoie un ou plusieurs signaux en forme d'impulsion sur le *latch* afin de le rendre passant. Ces architectures consomment un petit




**(a)** La plus simple expression de bascule à impulsion explicite [7].

**(b)** La bascule à impulsion explicite conventionnelle [40].

**Figure D.5:** Les structures basiques de bascules à impulsion explicite.




**(a)** Le concept de générateur d'impulsion double-front.

**(b)** GI double-front à portes de transmission NMOS [41].

**(c)** GI double-front à portes de transmission complètes [46].

**Figure D.6:** Générateurs d'impulsion (GIs) double-front.

plus que leurs pendants implicites. Néanmoins, elles ne nécessitent que de deux transistors minimum pour transmettre l'information de l'entrée à la sortie.

#### *D.1.2.4 Double-front*

Les bascules double-front, ou *dual-edge* en anglais, échantillonnent la donnée en entrée non pas sur un front - montant ou descendant - de l'horloge, mais sur les deux fronts. Comme conséquence directe, cela permet de diviser l'énergie dynamique de l'arbre d'horloge par deux, pour un même débit dans la logique combinatoire.

Chaque architecture des sections précédentes pourraient être adaptée en double-front. Néanmoins, [7] a montré que pour certaines topologies, la charge d'horloge est multipliée par deux ou plus, et donc le bénéfice d'une fréquence de commutation divisée par deux est annihilé. La littérature montre également que le moyen le plus efficace d'implémenter la fonction double-front est de modifier le générateur d'impulsion des bascules à impulsion explicite (voir figure D.6).

Par ailleurs, ce système de fonctionnement demande un rapport cyclique identique entre le niveau haut et le niveau bas de l'horloge, afin de ne pas dé-balancer la contrainte en temps du circuit. Or, cette contrainte est très difficilement atteignable à très basse tension d'alimentation, où la variabilité locale devient

prépondérante. Cette fonctionnalité sera donc difficilement implémentable pour les applications visées.

#### *D.1.2.5 Synthèse et impulsion explicite*

Le tableau D.1 résume la comparaison des quatre catégories de bascule.

**Table D.1:** Comparaison de topologies de bascules pour différentes figures de mérite.

| Topologie | Vitesse | Puissance | Surface | robustesse |
|---|---|---|---|---|
| Maitre-esclave | – | + | + | ++ |
| Différentielle | ++ | – – | – | – – |
| A impulsion implicite | + | – | – | – |
| A impulsion explicite† | ++/++ | -/++ | -/++ | -/- |
| Double-front | + | ++ | + | – |

† sans/avec générateur d'impulsion partagé

*Impulsion explicite*

Les architectures de bascule à impulsion explicite présentent plusieurs avantages que nous résumons ici :

- Elles sont rapides car ne présentant qu'un seul *latch* dans le chemin de propagation entrée-sortie,
- Elles ont néanmoins une consommation énergétique raisonnable car ne nécessitent pas de pré-charge à chaque cycle d'horloge,
- Par rapport aux impulsion implicites, elles sont plus rapides car demandant un empilement moins grand mais consomment un peu plus. Néanmoins, nous pouvons remarquer que le signal d'impulsion ($Pulse$ sur la figure D.5) peut être partagé sur plusieurs *latch* en parallèle, afin de diminuer la consommation par bascule.

Néanmoins, cette topologie présente deux grands désavantages :

Le temps de maintien est positif. Cela signifie que, lors de la synthèse et du placement et routage du circuit, les outils automatiques vont devoir insérer des tampons en délais sur les chemins courts afin de garantir la contrainte de maintien. Cela amènera une surconsommation en énergie et en surface.

Le deuxième problème, directement lié à l'architecture, est la variabilité du signal d'impulsion. En effet, le délais généré étant soumis à la variabilité locale comme tout signal, il est possible que le signal d'impulsion n'atteigne pas une tension suffisante pour permettre l'ouverture du *latch*. Cela signifie que, quelque

soit la fréquence de fonctionnement du circuit, la bascule ne sera jamais fonctionnelle.

Ces deux problèmes seront traités dans la section D.3 grâce à la technique du *current-starved*.

Dans la suite, la conception de bascules à impulsion explicite et haute efficacité énergétique en technologie FDSOI est étudiée. Comme il a été vu, l'étude peut être séparée entre le *latch* et le générateur d'impulsion, qui seront par conséquent le sujet des deux prochaines sections.

## D.2 COMPARAISON D'ARCHITECTURES

Après avoir parcouru l'état de l'art de la littérature, nous avons sélectionné six architectures qui répondent à nos contraintes. Ces dernières sont :

- Logique CMOS statique,
- Pas de chemin de court-circuit,
- Une porte de transmission à transistor unique est interdite,
- Un empilement de quatre transistors ou plus est aussi interdit,
- Les bascules sont sensibles sur un seul front d'horloge,
- La sortie des bascules doit être la sortie d'un inverseur,
- Toutes les entrées arrivent sur une grille de transistor,
- Une sortie inversée logiquement ou un front descendant actif sont autorisés,
- Les bascules possèdent les fonctions *scan* et *reset*.

La dernière contrainte provient du fait qu'aujourd'hui en industrie, et tout spécialement avec des nœuds technologiques avancés, il est nécessaire de tester le circuit quand il sort de fabrication. De plus, la fonction *reset* permet au concepteur de l'architecture du processeur de connaitre l'état du circuit à un moment donné.

Les six architectures sont représentées aux figures D.7 à D.10.

Les figures D.7 et D.8 représentent deux fois deux variantes de l'architecture conventionnelle.

Les versions TGPL et $C^2$MOS consistent en l'utilisation ou non d'une porte de transmission sur l'étage d'entrée. Une porte de transmission sera plus rapide qu'un inverseur trois-états. Néanmoins, les capacités de jonction des autres transistors de l'étage d'entrée seront toutes et complètement chargées et déchargées lors de chaque variation de la donnée en entrée. Il y a donc un compromis à faire entre la vitesse et la consommation.

Les versions -Data ou -Clk consistent en deux multiplexage différents des signaux de données ($D$ et la donnée de test $TI$). Soit l'étage d'entrée est un multiplexeur commandé par le signal de test $TE$, soit la génération d'impulsion est dédoublée. Dans le premier cas, nous voyons un empilement de trois transistors et dans le second cas, une augmentation de la capacité interne du générateur d'impulsion. C'est donc un deuxième compromis vitesse/consommation.

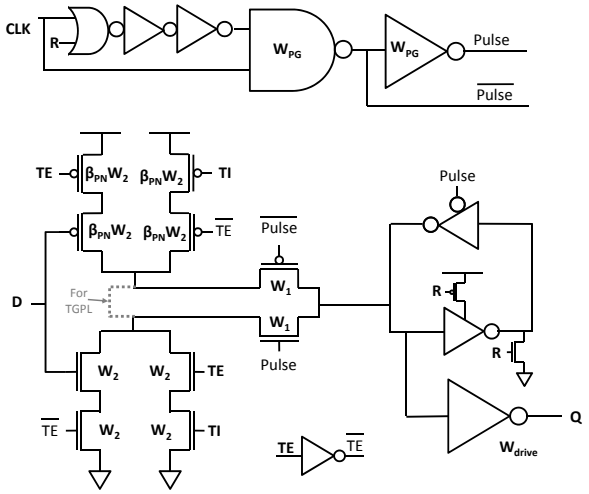



**Figure D.7:** Les bascules TGPL-Data and C$^2$MOS-Data (respectivement avec et sans les pointillés), deux variantes de l'architecture TGPL de la figure D.5.

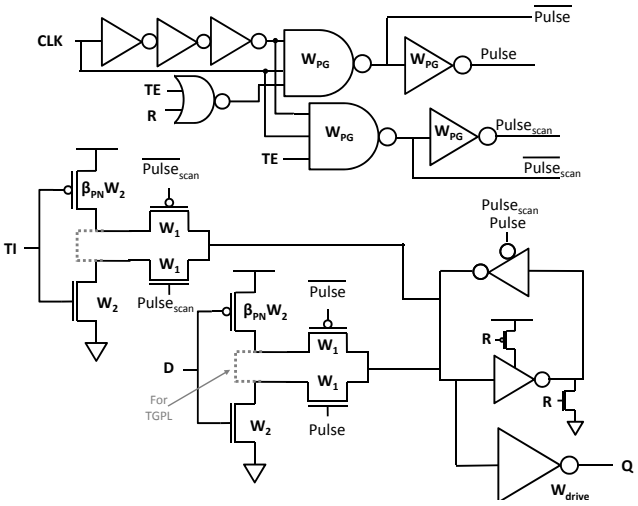



**Figure D.8:** Les bascules TGPL-Clk and C$^2$MOS-Clk (respectivement avec et sans les pointillés), deux variantes de l'architecture TGPL de la figure D.5.

L'architecture de la figure D.9 est une modification de l'architecture CDFF proposée dans [46], afin qu'elle répondent à nos critères précédemment cités. Nous pouvons noter que cette architecture n'a besoin que du signal d'impulsion $Pulse$, et non de son complémentaire logique.

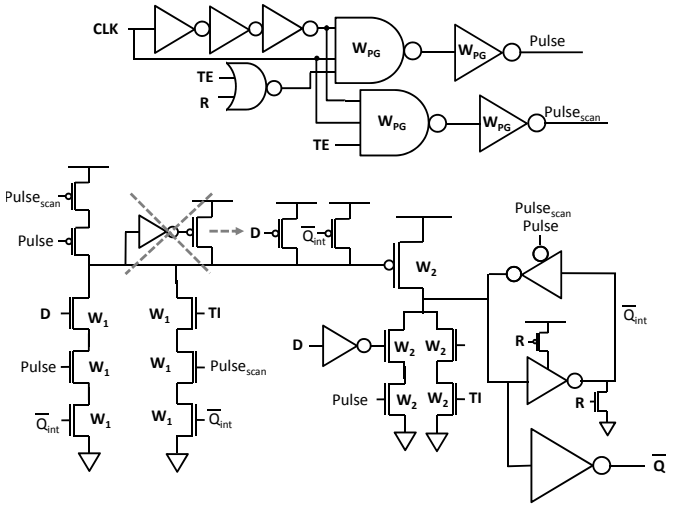



**Figure D.9:** La version avec *scan* et *reset* de l'architecture CDFF de la figure 1.17d

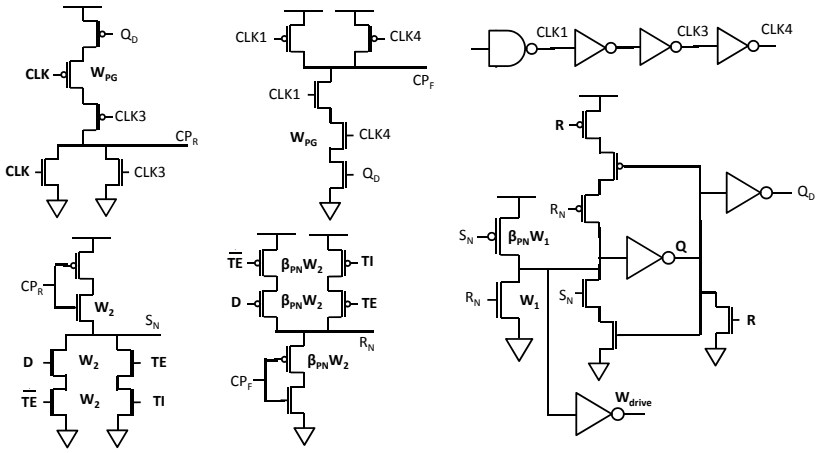



**Figure D.10:** CP$^3$L-Data, version avec *scan* et *reset* de l'architecture CP$^3$L de la figure 1.21.

De la même manière, la figure D.10 représente une autre version de l'architecture CP$^3$L proposée dans [49].

Dans la suite, nous allons comparer ces architectures dans l'espace énergie-délais, grâce à leur courbe d'efficacité énergétique (CEE).

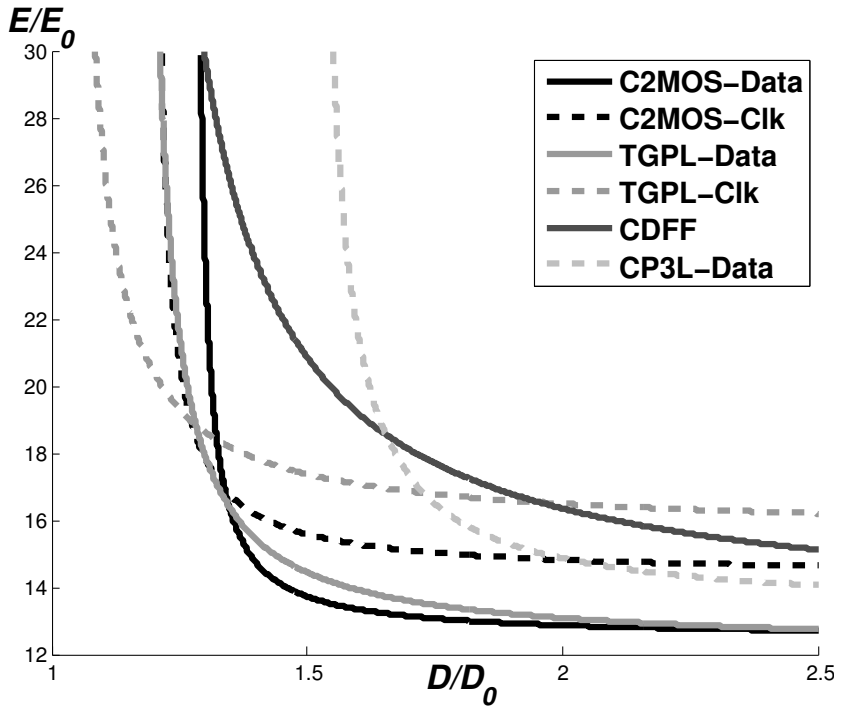



**Figure D.11:** CEE: $V_{dd}$ = 1.0V, Vdds=Gnds=0V, $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, corner TT, température 70°C.

### D.2.1 Comparaison dans le plan énergie-délais

Ces six architectures ont été comparées dans le domaine énergie-délais ($E - D$) grce à un algorithme de dimensionnement, selon la méthode présentée dans [6]. Cette méthode consiste à déterminer les points de dimensionnement qui donnent une valeur optimale à une figure de mérite donnée. A partir de ces points, nous pouvons inter- et extrapoler la courbe d'efficacité énergétique (CEE), présentant une forme d'hyperbole [66]. Il est à noter que la principale différence entre notre banc de test et celui utilisé dans [6], est l'utilisation d'une métrique différente pour le temps de *setup*. Cette dernière est basée sur le temps de transition de la sortie et convient parfaitement pour les bascules à impulsion [28].

Comme nous visons la large gamme de tension, les architectures de bascules à impulsion ont été comparées à tension d'alimentation nominale et à très faible tension, pour des tensions de substrat nominales, soit Gnds= Vdds = 0V.

#### *D.2.1.1 Tensions de bias nominales*

Les courbes d'efficacité énergétique de toutes les architectures de bascules sont représentées à la figure D.11, pour une tension d'alimentation nominale. La

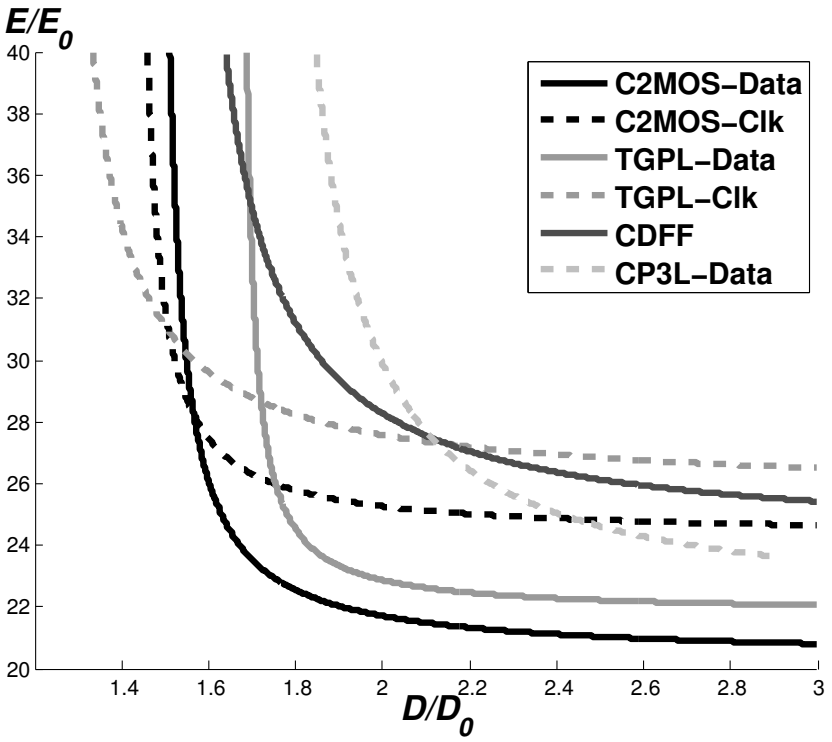



**Figure D.12:** CEE: $V_{dd}$ = 0.35V, Vdds=Gnds=0V, $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, corner TT, température 25°C.

température est de 70°C, le facteur d'activité ($\alpha_{sw}$) est de 15% and la période d'horloge ($T_{clk}$) est 40 fois le délais de propagation d'une chaine d'inverseur de *fanout* 4 dans les mêmes conditions environnementales. Toutes ces valeurs sont typiques pour les applications visées et ne changent pas profondément le résultat de comparaison [6].

Nous voyons que les architectures *-Data produisent la plus faible consommation d'énergie, avec la C$^2$MOS légèrement meilleure que la TGPL. Durant une transition de la C$^2$MOS, les nœuds intermédiaires de l'inverseur trois-états d'entrée ne sont pas complètement chargés ou déchargés, parce que la tension grille-source du premier transistor de l'empilement diminue graduellement en dessous de la tension de seuil. Par conséquent, l'architecture C$^2$MOS sauvegarde de l'énergie dynamique dans la charge et décharge des capacités de jonction proportionnelles à $W_2$ (voir figure D.7). Cette consommation dynamique plus faible permet aux architectures C$^2$MOS d'atteindre un plus grand $W_k$, donc une plus grande vitesse, pour la même énergie. Néanmoins, les C$^2$MOS sont finalement dépassées par les topologies TGPL dans la région haute performance, car

les deux transistors de la porte de transmission aident à l'augmentation de la vitesse. Les architectures *-Clk produisent un meilleur compromis $E - D$ dans la région haute performances, grce à leur empilement d'entrée plus faible. Enfin, notons que les deux dernières structures présentent un compromis beaucoup moins intéressant sur tout l'espace $E - D$.

Les courbes d'efficacité énergétique extraites à très faible tension d'alimentation - 0.35V, fournissant l'énergie par opération minimum pour la technologie 28nm FDSOI - sont représentées à la figure D.12. Ici la température est de 25°C car l'auto-échauffement des circuits à très basse consommation est pratiquement négligeable, et ces circuits fonctionnent normalement à température ambiante.

Nous voyons que la courbe C$^2$MOS-Data est la plus efficace en énergie sur presque tout le plan $E - D$, excepté dans la région de très haute performance. Grce à son empilement, cette architecture exhibe un très faible courant de fuite, qui devient extrêmement significatif à très basse tension et donc impacte fortement la consommation.

À très faible tension d'alimentation, les architectures de type TGPL sont moins efficaces en énergie sur tout le plan $E-D$. En plus de présenter un courant de fuite plus de grand que les architectures de type C$^2$MOS, leur bénéfice en vitesse est mis en péril par le régime sous seuil. En effet, durant une transition, la tension grille-source $V_{GS}$ de l'un des transistors de la porte de transmission diminue progressivement. Comme il dépend exponentiellement de $V_{GS}$ dans le régime sous-seuil, le courant de drain de l'un des transistors est très vite déjà négligeable après le début de la transition. Cela mène à une architecture de type C$^2$MOS, avec de plus capacités de jonction dans le chemin entrée-sortie.

A nouveau, les deux dernières structures présentent un compromis beaucoup moins intéressant sur tout l'espace $E-D$. Le fait que la structure CP$^3$L présente un moins bon produit énergie-délais (PED), en opposition aux résultats de [49], montre que les fonctionnalités de *reset* et surtout de *scan* peuvent changer les conclusions de la comparaison. Ainsi, cela prouve que la facilité d'implémenter la fonction *scan* doit être prise en compte dans le choix de l'architecture de la bascule.

#### *D.2.1.2 Modification de la tension de bias*

Dans la section précédente, nous avons vu que pour être le plus efficace en énergie sur tout le plan $E - D$, nous avons besoin de plusieurs architectures et surtout de plusieurs points de dimensionnement. Toutes ces architectures dimensionnées devront être caractérisées pour plusieurs conditions PVT, et cela pourrait significativement augmenter le temps de conception.

Cependant, ce compromis entre temps de développement et efficacité énergétique peut être presque complètement éviter grce à la technologie FD-SOI. Les figures D.13 et D.14 comparent les performances dans l'espace $E - D$ de l'architecture C$^2$MOS-Data (dimensionnement minimum PED) sur laquelle est appliquée une large et symétrique polarisation arrière. Comme nous pouvons voir, le délais dans la région haut-performance est plus petit, pour la même

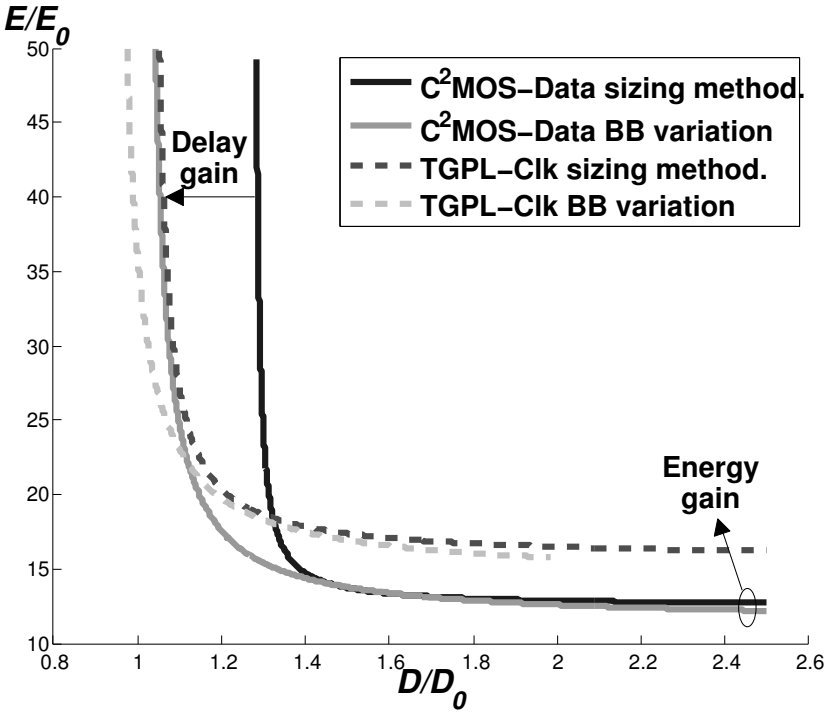



**Figure D.13:** CEEs des architectures C$^2$MOS-Data et TGPL-Clk extraites partir de l'algorithme de dimensionnement et partir d'une polarisation face arrire ($V_{dd}$ = 1V, Vdds/Gnds range = $\pm$ 1V , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, température 70°C). La technique de polarisation face arrire applique sur une large gamme fournie de meilleures performances en dlais et nergie que la mthodologie de dimensionnement.

consommation énergétique, que le délais obtenu par la méthode de dimensionnement. De la même façon dans la région basse-puissance, la consommation est plus petite pour le même délais que l'énergie consommée par le dimensionnement minimum des transistors.

À très faible tension, l'impacte de la tension de seuil sur le courant en mode passant, est presque aussi important que sur le courant de fuite en mode bloqué. Par conséquent, une augmentation (diminution) du délais entraine une diminution (augmentation) du courant de fuite. L'énergie de fuite étant le courant de fuite intégré sur une période d'horloge, l'énergie par opération ne varie pas de manière importante dans la région basse-puissance, *i.e.* pour une polarisation arrière en inverse, et donc reste relativement la même que celle à tension de face arrière nominale. D'un autre côté, le gain en délais est relativement bien plus grand à très faible tension qu'à tension d'alimentation nominale. Tout cela montrent clairement que nous pouvons moduler les performances en énergie et délais

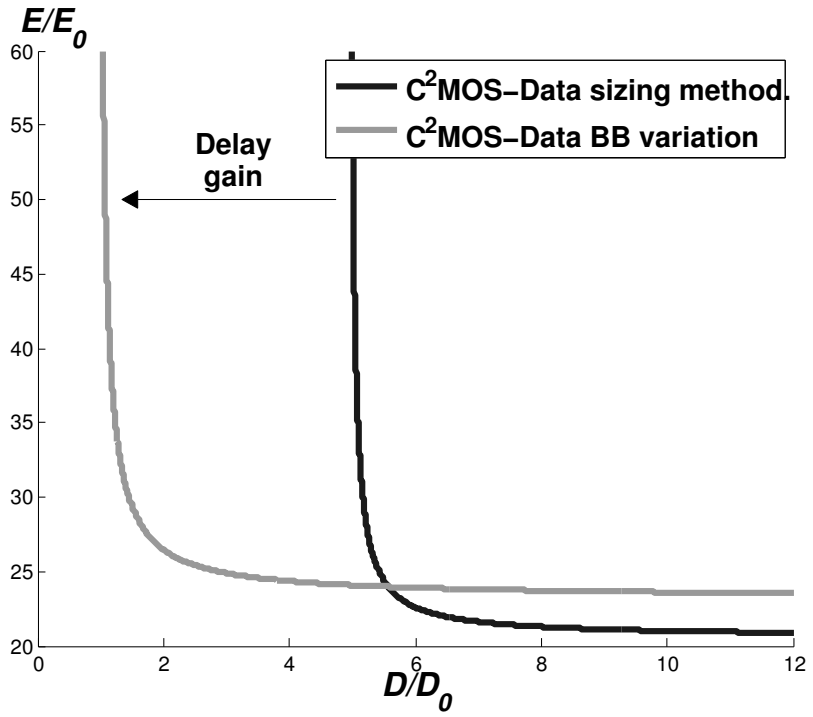



**Figure D.14:** CEEs de l'architecture C$^2$MOS-Data extraites partir de l'algorithme de dimensionnement et partir d'une polarisation face arrire ($V_{dd} = 0.35$V , $\alpha_{sw} = 0.15$, $T_{clk} = 40FO4$, température 25°C).

plus efficacement que par la méthodologie de dimensionnement, et de plus, de manière dynamique durant la vie du circuit.

### D.2.2 Intégration dans le flot et résultats de mesure

Une architecture de bascule à impulsion explicite a été dessinée, caractérisée et incorporée dans le flot de conception du circuit FRISBEE [24]. Il faut tout d'abord noté, qu'aucun problème additionnel majeur ne fut rencontré durant ces étapes. Les outils automatiques ont correctement interprété nos bascules à impulsion, pourtant fort différentes des bascules maitre-esclave classiques. Ensuite, il est à souligner que les mesures de FRISBEE ont montré des performances en vitesse et en consommation qui surpassaient, comme attendu, celles de l'état de l'art.

En parallèle de ça, des architectures de bascule isolées ont été fondues sur silicium afin d'étudier plus précisément les caractéristiques des bascules elles-mêmes. Pour chaque architecture, nous avons pu mesuré le délais de propagation de 63

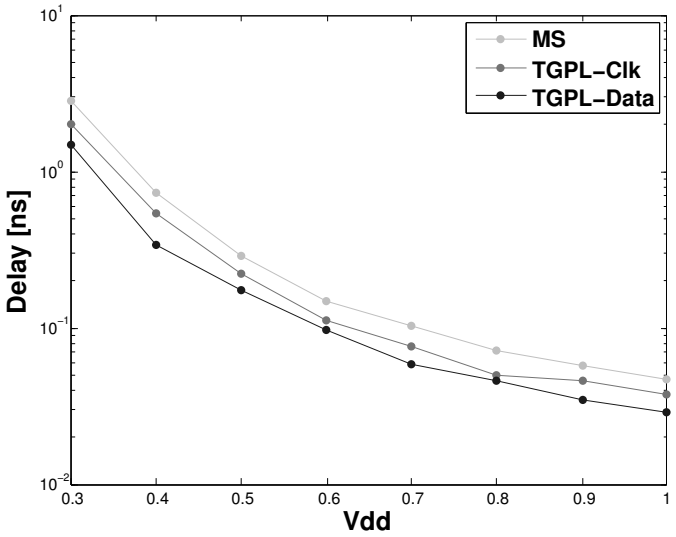


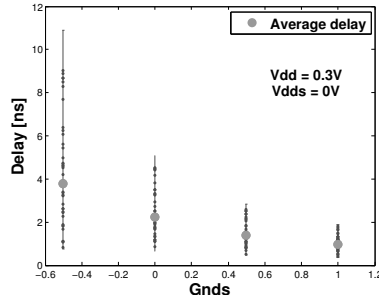

**Figure D.15:** Délais mesurés de 0.3V à 1V dans le cas nominal (Vdds=Gnds=0V). Le délais Clk-Q est plus faible pour la TGPL-Data que pour la TGPL-Clk, contrairement au délais entrée-sortie (T°= 80°C).

bascules, réparties uniformément sur la galette de silicium, afin d'en extraire une moyenne statistique.

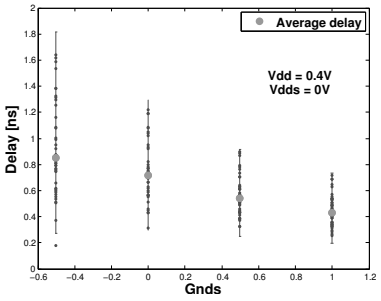
La figure D.15 montre le délais horloge-sortie (Clk-Q) moyen mesuré pour différentes architectures à plusieurs tension d'alimentation. Comme prédit en simulation, l'architecture TGPL-Clk a un plus grand Clk-Q que la TGPL-Data mais néanmoins un plus faible délais de propagation entrée-sortie (D-Q). Dans tous les cas, les bascules à impulsion sont bien plus rapides que la bascule maitre-esclave.

La figure D.16 représente les délais Clk-Q mesurés à très basse tension d'alimentation, pour plusieurs configuration de tensions de substrat. Nous voyons très clairement que, grce à la technologie FDSOI, nous pouvons faire varier les performances d'un facteur très significatif à très basse tension (figures D.16a et D.16d). De plus, nous voyons également pour les cas $V_{dd} = 0.3V$ que la variation de la tension de substrat induit un changement de régime, *i.e.* en-dessous et au-dessus de la tension de seuil du transistor.

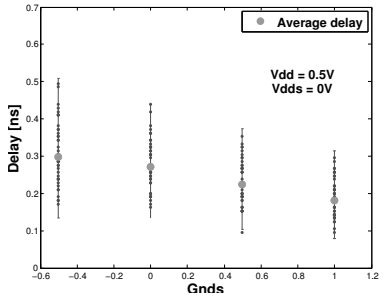
Finalement, le tableau D.2 et la figure D.17 comparent l'efficacité énergétique des bascules à impulsion et maitre-esclave, en incluant l'énergie par simulation *post-layout*. Nous voyons que nos bascules surpassent effectivement les bascules maitre-esclave (ME) dès que l'on tient compte du délais de propagation - pour le produit énergie-délais par example (PED). Néanmoins, il faut que les bascules



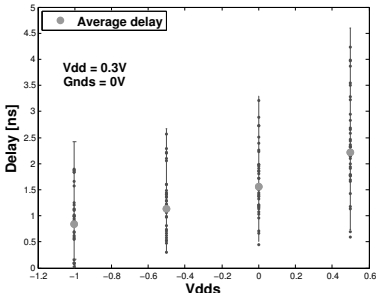

**(a)** Gnds variation à tension d'alimentation sous-seuil.




**(b)** Gnds variation à tension d'alimentation proche du seuil.




**(c)** Gnds variation à tension d'alimentation au-dessus du seuil.



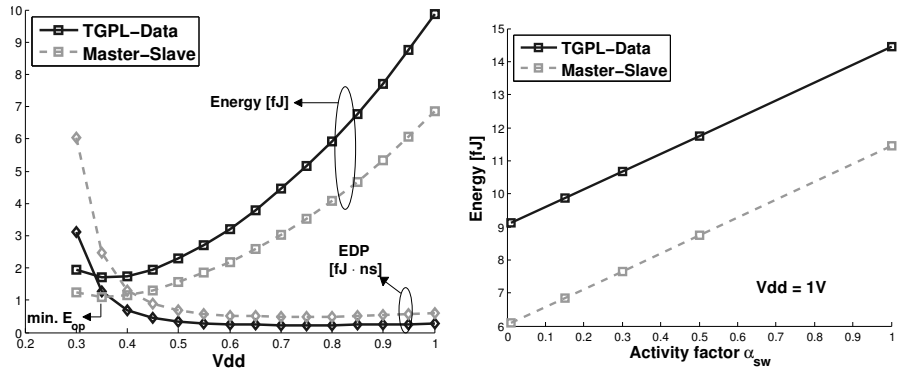

**(d)** Vdds variation à tension d'alimentation sous-seuil.

**Figure D.16:** Évolution du délais mesuré avec Vdds et Gnds.

ME présentent toujours une plus faible consommation énergétique, quels que soient la tension d'alimentation et le facteur d'activité. Nous y reviendrons dans la section D.4.3

**Table D.2:** Comparaison des figures de mérite PED et PED$^2$ pour les trois architectures (simulations *post-layout* à 1V et 80°C).

| Architecture | $E_{op}[fJ]$ ($\alpha_{sw} = 15\%$) | PED [fJ $\cdot$ ps] | PED$^2$ [fJ $\cdot$ ps$^2$] | Surface [$\mu m^2$] |
|---|---|---|---|---|
| MS | 6.72 (ref.) | 1136 (ref.) | 1877 | 4.4 (ref.) |
| TGPL-Data | 10.08 (+50%) | 288 (-74%) | 82 | 5.4 (+23%) |
| TGPL-Clk | 14.88 (+121%) | 334 (-70%) | 74 | 6.7 (+52%) |




**(a)** Évolution de l'énergie moyenne par cycle ($\alpha_{sw} = 15\%$) et du produit énergie-délais (PED) avec la tension d'alimentation.

**(b)** Évolution de l'énergie moyenne par cycle avec le facteur d'activité.

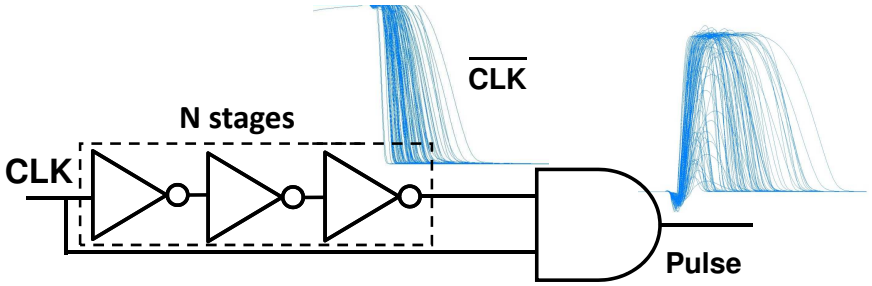
**Figure D.17:** L'architecture TGPL surpasse la topologie maitre-esclave pour toute tension d'alimentation entre 0.3V et 1V et tous les facteurs d'activité (simulation *post-layout*).

## D.3 GÉNÉRATEUR D'IMPULSION ROBUSTE

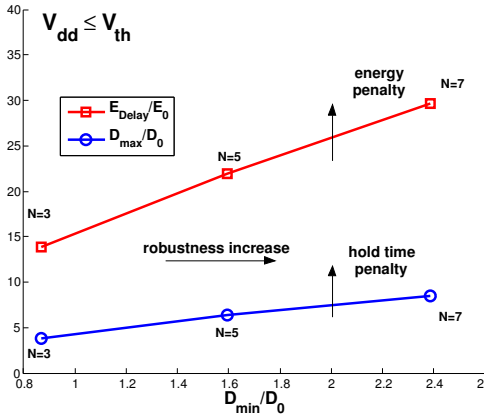
Pour rappel, les bascules à impulsion sont composées d'un *latch* ouvert durant une courte période après le front d'horloge déclencheur. Cette période est physiquement déterminée par un signal d'impulsion généré par le générateur d'impulsion (GI). La largeur de cette impulsion est donnée par le délais entre l'horloge et la sortie du générateur de délais (voir figure D.18a).

À très basse tension d'alimentation, où l'impact des variations locales est prédominant, et le délais généré et le délais entrée-sortie ($D - to - Q$)varient très significativement d'une bascule à l'autre sur le circuit. Or, avec un lent chemin $D - to - Q$ dans la bascule, le signal d'impulsion pourrait être trop étroit pour permettre un échantillonnage de la nouvelle donnée. Une solution basique et classique est d'ajouter des étages dans la chaine de délais du générateur d'impulsion (figure D.18a). Néanmoins, cela augmente dramatiquement le temps de maintien, défini par la largeur maximale d'impulsion obtenue par le GI soumis aux variations locales. Si le temps de maintien est bien plus grand que le délais de propagation $D - to - Q$, beaucoup de tampons en délais seront insérés sur les chemins courts.

En conséquence, il y a une consommation d'énergie supplémentaire venant de l'insertion et d'inverseur dans le GI, et de tampons en délais sur les chemin courts pour respecter la contrainte de temps de maintien. Comme montré à la figure D.18b, il y a un compromis entre la robustesse de la bascule et la consommation énergétique.




**(a)** Problèmes de variabilité dans les générateurs d'impulsion à très basse tension d'alimentation.




**(b)** La dissipation énergétique et le délais maximum ($\mu + 3\sigma$) en fonction du délais minimum ($\mu - 3\sigma$) d'une chaine d'inverseurs (N = 3, 5 and 7)

**Figure D.18:** Augmenter la robustesse d'une bascule à impulsion conduit à un grand coût énergétique.

Voilà pourquoi, j'ai proposé un générateur de délais (GD) basé sur la technique du *current-starved* [72]. L'architecture est présentée à la figure D.19. Notre GD présent seulement trois étages dans le chemin de délais, mais augmente le délais moyen grce aux transistors *current-starved* placé entre les inverseurs et la leur tension d'alimentation. Comme ces transistors toujours ouverts ne sont pas dans le chemin de délais, ils n'impactent pratiquement pas la dispersion du délais. Ainsi, cette architecture *current-starved* peut être vue comme une translation sans expansion de la fenêtre de délais sur la ligne du temps. Cette translation sans expansion est le point clé permettant de garantir une robustesse suffisante sans surconsommation énergétique.

Insistons encore une fois sur le fait que les tailles des transistors *current-starved* sont des degrés de liberté réellement puissants et faciles à manier. Finale-

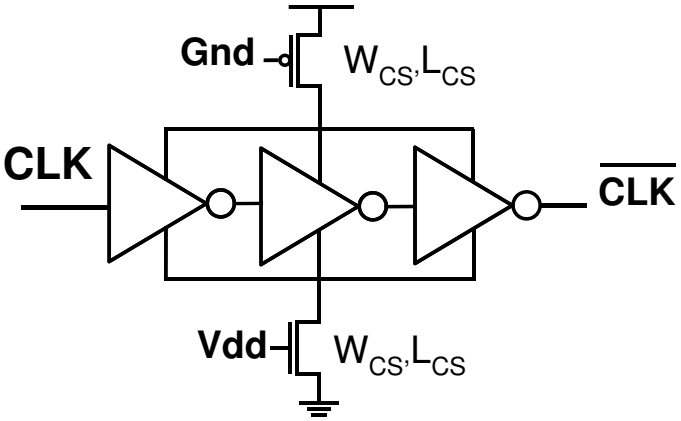



**Figure D.19:** Architecture du générateur de délais (GD) proposé [70].

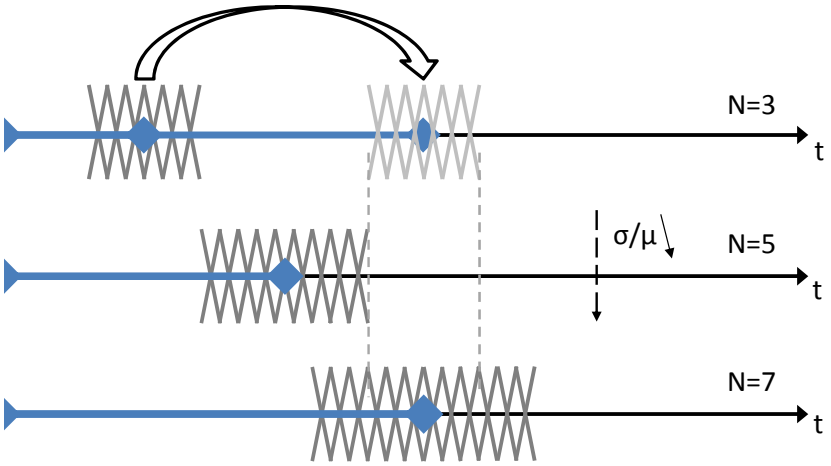



**Figure D.20:** L'architecture de GD dite *current-starved* effectue une translation sans expansion de la fenêtre de délais sur la ligne du temps. En d'autres mots, elle augmente le délais moyen sans augmentation notable de la dispersion.

ment, notons que cette technique peut facilement se transposer aux générateurs d'impulsion sur double fronts [73, 74].

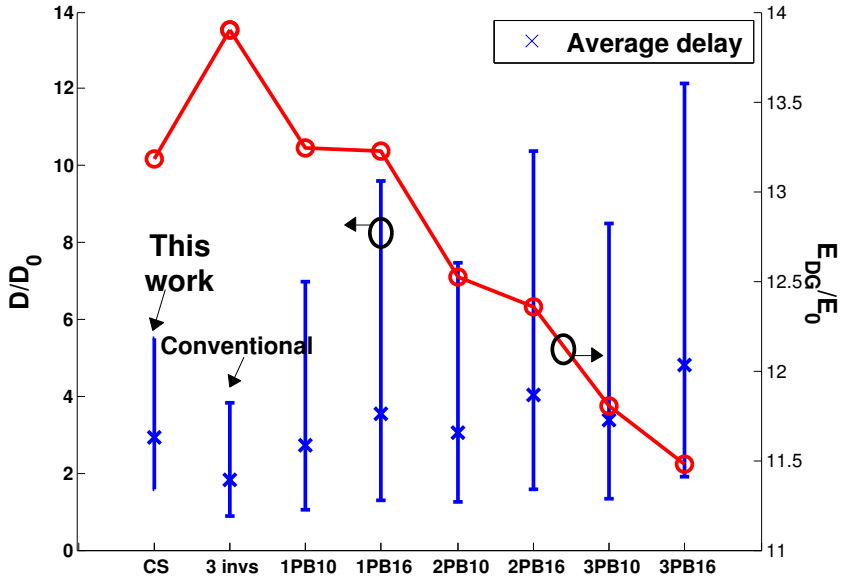



**Figure D.21:** Comparaison des performances nergtiques et en dlais dans le rgime sous-seuil ($V_{dd} = 0.3V$). Le nom indique l'augmentation de la longueur de polysilicium and le nombre d'tages prsentant celle-ci.

### D.3.1 Comparaison de générateur de délais

Les figures D.21 et D.22 comparent le compromis énergie-robustesse entre plusieurs architecture de GD de l'état de l'art et notre architecture *current-starved* avec un dimensionnement donné. Chaque architecture présente un plus grand délais minimum, càd robustesse, que l'architecture conventionnelle. Néanmoins, c'est notre architecture de GD qui exhibe le plus petit délais maximum, càd consommation, après l'architecture basique.

Ensuite alors, nous avons comparé la robustesse de notre GD *current-starved* et le GD classique grâce à des simulations *post-layout*. Comme montré au tableau D.3, nous obtenons une amélioration substantielle de la robustesse de la bascule grâce à notre architecture de générateur de délais.

Afin de démontrer encore plus cette amélioration, la section suivante présente des résultats silicium qui corroborent ces dires.

### D.3.2 Robustesse : mesures silicium

Une puce de test a été fabriquée en technologie 28nm FDSOI pour tester la fonctionnalité d'une bascule pour des paramètres vdd, vdds et gnds donnés. À nouveau, 64 bascules réparties uniformément sur la galette de silicium ont pu être mesurées afin d'étudier statistiquement les tensions extrêmes de fonctionnement.

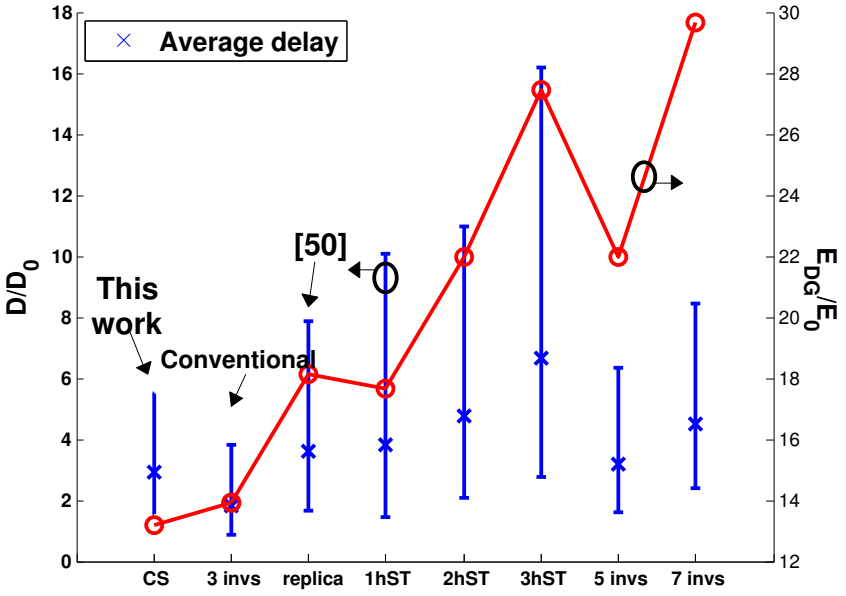


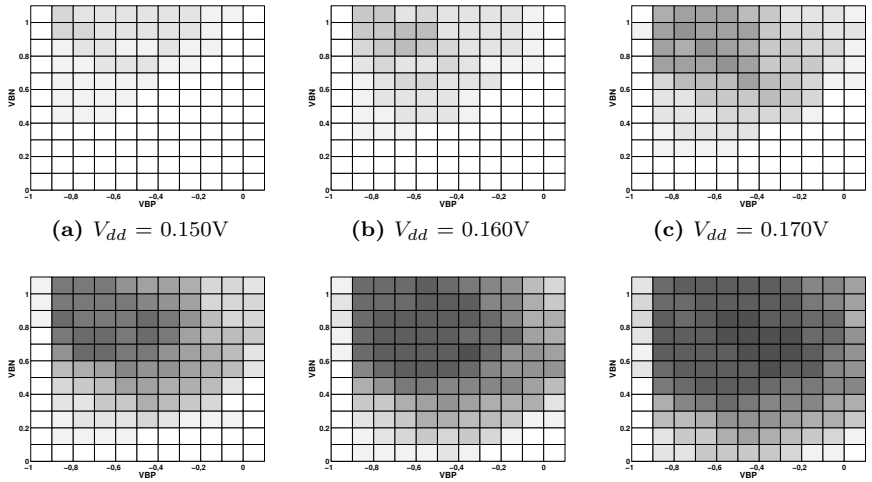

**Figure D.22:** Comparaison des performances nergtiques et en dlais dans le rgime sous-seuil ($V_{dd} = 0.3V$). pour diffrentes architectures de GD.

**Table D.3:** Comparaison de la robustesse de GDs avec 1000 tirages Monte-Carlo de simulations post-layout en rgime sous-seuil (0.3V)

| Conditions PVT | # Échecs | |
|---|---|---|
| | Conventionnel | Propos |
| FS -40° | 14 | 0 |
| FS 85° | 3 | 0 |
| SF -40° | 60 | 1 |
| SF 85° | 29 | 0 |

Le rendement fonctionnel est représenté à la figure D.23 dans le plan (Vdds,Gnds). Nous voyons que le rendement maximum est obtenu pour un ou plusieurs points le long de la diagonale 1 (voir figure D.24) et qu'il diminue ensuite progressivement vers les quatre coins du plan. La distribution du rendement est expliquée de manière approfondie dans CP.6 (voir liste des publications).

Comme montré à la figure D.24, la décroissance le long de la diagonale 2 s'explique par le débalancement entre les transistors NMOS et PMOS [77]. En effet, les tensions de seuil de ces deux transistors évoluent dans des directions opposées le long de cette diagonal. Il en va donc de même pour le rapport des courants en état passant et bloqué.




**(a)** $V_{dd} = 0.150\text{V}$ **(b)** $V_{dd} = 0.160\text{V}$ **(c)** $V_{dd} = 0.170\text{V}$

**(d)** $V_{dd} = 0.180\text{V}$ **(e)** $V_{dd} = 0.190\text{V}$ **(f)** $V_{dd} = 0.200\text{V}$

**Figure D.23:** Rendement mesuré de l'architecture TGPL-Data dans l'espace (Vdds,Gnds) à 25°C. Plus le carré est sombre, plus le rendement est élevé.

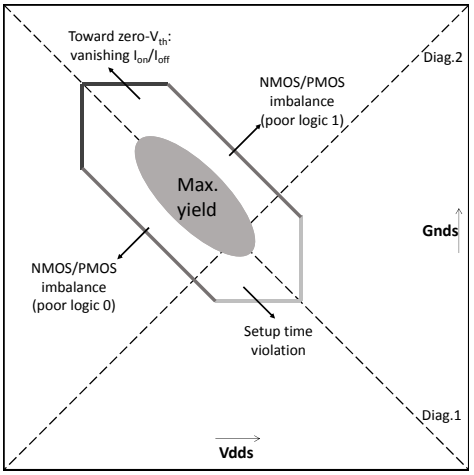



**Figure D.24:** Comportement schématique du rendement dans l'espace Vdds-Gnds.

La décroissance du rendement vers le coin supérieur gauche s'explique également par la diminution du rapport des courants, mais pour une raison différente. Le long de la diagonal 1, les tension de substrat sont symétriques et donc les courants évoluent dans la même direction. Néanmoins, diminuer la tension de seuil des transistors peut mener à un transistor avec une tension de

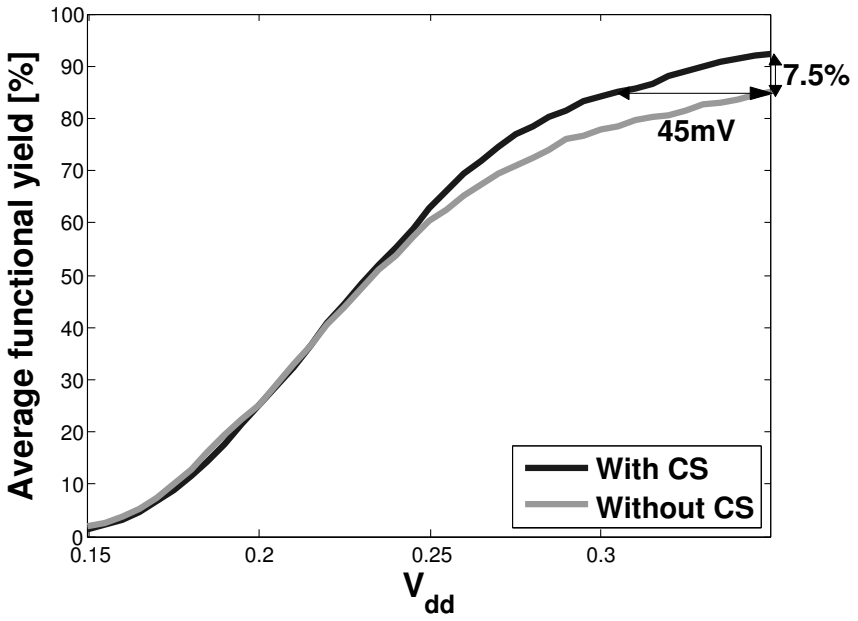



**Figure D.25:** The current-starved (CS) delay generator provides a notable gain for the yield.

seuil de 0V. Par conséquent, une polarisation arrière vers l'avant augmente substantiellement la vitesse des circuits à très basse tension (voir section D.2.1.2) mais diminue en même temps la robustesse de ces circuits.

Finalement, le bord en gris clair sur la figure D.24 est dû à la violation du temps de *setup*. La période d'horloge pour notre banc de mesure était de 2000ns et, à très faible tension avec des polarisations arrières nominales, les bascules sont trop lentes pour changer d'état.

### *Comparaison de rendement*

A partir des mesures, nous pouvons calculer la moyenne du rendement sur tout le plan (Vdds,Gnds) en fonction de la tension d'alimentation. La figure D.25 compare ces courbes pour deux architectures de bascules : l'une avec un générateur de délais conventionnel, et l'autre avec la technique du *current-starved*. En comparant aux autres architectures, nous voyons que le GD *current-starved* permet d'atteindre le même rendement fonctionnel à une tension d'alimentation jusqu'à 45mV inférieur et un rendement 7.5% supérieur pour une même tension d'alimentation.

Nous pouvons par conséquent affirmer que, notre GD avec la technique *current-starved* augmente significativement la robustesse des bascules à impulsion.

## D.4 RÉDUCTION DE LA CONSOMMATION D'ÉNERGIE

Dans cette quatrième et dernière section est présentées des techniques architecturales permettant de réduire la consommation des architectures de bascule impulsion explicite.

### D.4.1 Technique de capture conditionnelle

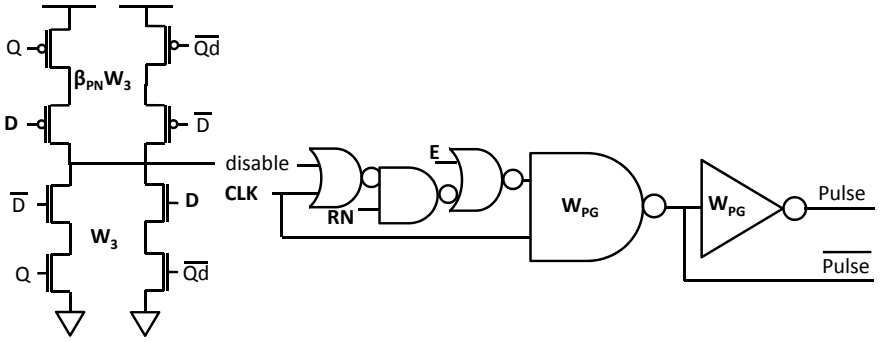
Jusqu'à présent, nous avons clairement vu que les bascules à impulsion explicite sont bien plus rapides que les bascules de type maitre-esclave (ME), bien qu'elles présentent de plus grandes consommation et surface. Après avoir garantit la robustesse de nos bascules dans la section D.3, ce chapitre étudie plusieurs techniques pour diminuer leur consommation et surface des bascules et finalement atteindre des performances globales meilleurs que les MEs.

Tout d'abord, nous identifions par simulation que la principale source de dissipation d'énergie est le générateur d'impulsion (GI). En effet, les cinq étages du GI (3 inverseurs de tailles minimale, une porte *NAND* et un inverseur) sont activés à chaque cycle d'horloge, même sans activité sur la donnée en entrée. Pour s'attaquer à cette dissipation inutile d'énergie, l'architecture $C^2$MOS-Xor, représentée à la figure D.26, est proposée. Cette architecture utilise une technique de capture conditionnelle utilisant une pseudo porte XOR, avec une approche légèrement différente de celle de [80]. En utilisant la propriété d'impulsion explicite, une porte XOR est ajoutée à l'entrée du GI afin de désactiver la génération de l'impulsion quand à la fois l'entrée (D) et la sortie courante (Q) sont identiques, càd quand la capture n'est pas nécessaire.

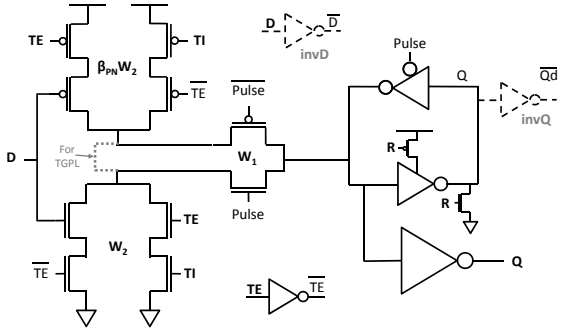
Notons que, maintenant, le temps de setup et de maintien sont, respectivement, positif et négatif. En effet, la donnée doit maintenant être valide suffisamment tt avant le front déclencheur de l'horloge tel que le signal *disable* rende le GI actif. Cela signifie que cette architecture résout le problème de temps de maintien positif des bascules à impulsion.

### D.4.2 Partage du générateur d'impulsion

Le partage du générateur d'impulsion (GI) a été étudié un certain nombre de fois d'un point de vue de l'efficacité énergétique [81, 82, 83, 84]. L'objectif de partager un GI avec N *latches* est d'obtenir une plus faible consommation énergétique que utiliser 2N *latches* avec une topologie maitre-esclave. La plupart des papiers dans la littérature travaillent dans le régime bien au-dessus du seuil, où les variations aléatoires peuvent être plus facilement traitées pour garantir une pente suffisante du signal d'impulsion après l'étape du placement et routage. Dans les régimes




**(a)** Le générateur d'impulsion conditionnelle.




**(b)** Le latch C$^2$MOS-Xor avec des inverseurs pour les signaux $\overline{D}$ et $\overline{Q_d}$.

**Figure D.26:** Schématique des architectures TGPL/C$^2$MOS-invQ/invD.

proches du et sous le seuil des transistors, nous considérons que le variabilité du signal d'impulsion est trop grand (voir section D.3) pour travailler avec une capacité de sortie du GI variable. Par conséquent, nous recommandons un unique bloc composé de plusieurs *latches* et d'un seul générateur d'impulsion, qui peut être caractérisé indépendamment.

Sur les figures D.27 et D.28, nous voyons qu'après un certain nombre de bascules (8 pour notre technologie), nous consommons moins d'énergie et occupons moins de surface par bascule. Néanmoins, nous voyons qu'à partir de 16 *latches*, le gain en énergie devient négligeable et seule la surface pourra motiver une augmentation du nombre de *latches* partagés.

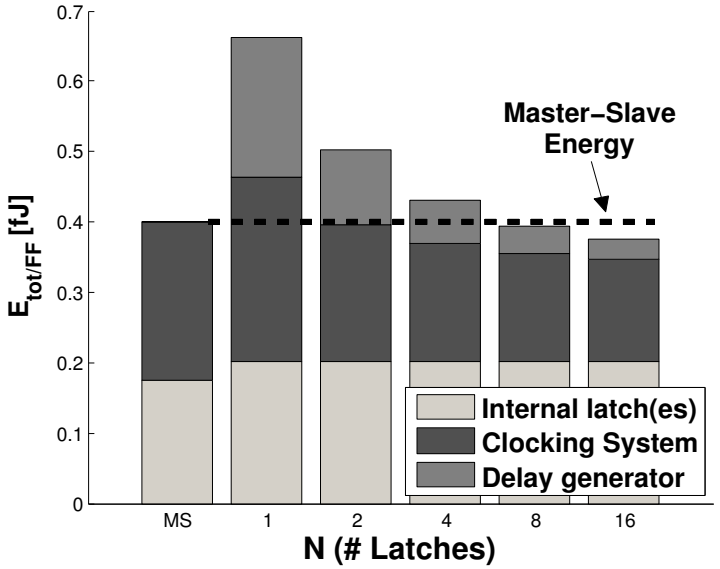



**Figure D.27:** L'énergie par bascule fonction du nombre de *latches* partageant un GI. Les barres représentent les proportions de l'énergie consummée dans le générateur de délais, la partie "horlogée" (les signaux $CLK$ et $Pulse$) et le *latch* pris isolément. $V_{dd}$=1V, $T_{clk}$ = 40FO4, $\alpha_{sw}$ = 15%.

### D.4.3 Banc de registres

Dans cette section, nous comparons deux bancs de registres (BR) : l'un avec nos bascules à impulsion et l'autre implémenté à partir de bascule maitre-esclave conventionnelles. Un banc de registres est une petite mémoire embarquée, synthétisée et inclue directement dans la logic combinatoire. C'est un élément primordial de chaque architecture de microprocesseur, von Neumann ou Harvard, représentant une part non-négligeable de la consommation énergétique de ce dernier [87]. Pour de faible taille de mémoire, les bancs de registres sont plus efficaces en énergie et en surface qu'une mémoire SRAM, et plus rapides dans tous les cas. De plus, ils sont beaucoup plus robustes à une diminution de la tension d'alimentation et peuvent être facilement intégrés dans la logique. En effet, la cellule fondamentale du banc de registres est la bascule.

Un banc de registres structurés a été dessiné, incorporant plusieurs innovations des sections précédentes : une architecture $C^2$MOS-Data avec un dimensionnement donnant le PED minimum, un générateur de délais comportant la

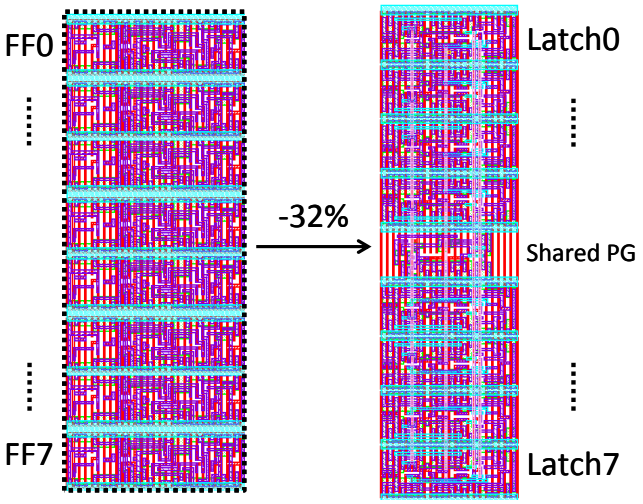
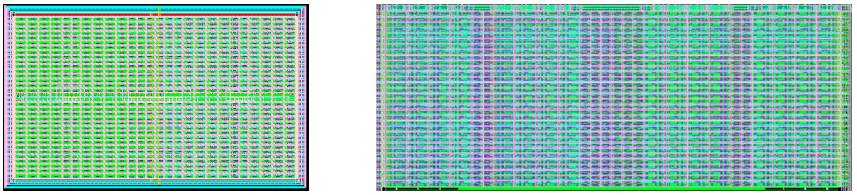



**Figure D.28:** Le bloc de 8 bascules à impulsion (droite) a 32% de surface en moins que les 8 bascules maitre-esclave.



**(a)** *Layout* du banc de registre structuré basé sur des bascule à impulsion.

**(b)** *Layout* du banc de registre structuré basé sur des bascule à maitre-esclave.

**Figure D.29:** La comparaison des *layouts* mène à un gain en surface de 14%.

technique *current-starved*, et un GI partagé pour chaque registre. Afin de mener l'étude dans des conditions réalistes, nous avons choisi les mêmes caractéristiques que le banc de registre du microprocesseur Cortex-M0 dédié à la très faible consommation : 16 registres de 32 *bits*, un port d'écriture synchrone et deux ports de lecture asynchrones.

Tout d'abord, la figure D.29 montre le *layout* des deux banc de registres, l'un avec les bascules à impulsion et l'autre avec les bascules maitre-esclaves (MEs). Le banc de registre composé de nos bascules à impulsion présente une surface 14% plus petite que celui basé sur les MEs.

La figure D.30 compare l'énergie par opération entre les deux bancs de registres (BR). Pour des tensions d'alimentation plus hautes que 0.5V, l'énergie par

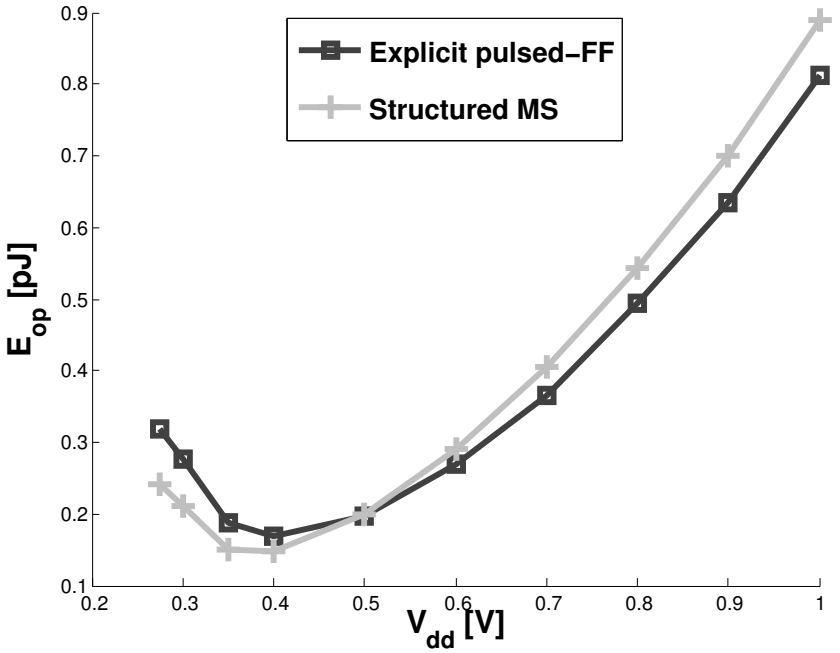



**Figure D.30:** Energie par opération sur une large gamme de tension (Gnds = Vdds = 0V et $\alpha_{rate} = 15\%$). Le banc de registres composé de bascules à impulsion explicite présente une plus faible consommation énergétique pour des $V_{dd}$ au-dessus du seuil, et une optimale à 0.35V.

opération moyenne $E_{op}$ est plus petite pour le BR contenant les bascules à impulsion. Ensuite, leur plus grand courant de fuite mène à une plus petite $E_{op}$ pour le BR contenant les MEs. En effet, à faible tension, la puissance statique de fuite représente une part de plus en plus important de l'énergie totale, et donc le courant de fuite pénalise notre structure de BR basé sur les bascules à impulsion.

## D.5 CONCLUSIONS

Dans ce travail, des architectures de bascules à impulsion robuste et efficace en énergie, visant la très large gamme de tension d'alimentation et les circuits très basse consommation, ont été développées et conçues en technologie FDSOI.

En premier lieu, la topologie de bascule à impulsion explicite a été mise en évidence parmi la littérature scientifique au chapitre 1. Cette architecture présente de remarquables propriétés temporelles, càd petit délais entrée-sortie, temps de *setup* négatif, facilités d'implémentation des techniques de *time-*

*borrowing* et de double-front, ainsi qu'un générateur d'impulsion partageable. Néanmoins, cette topologie est rarement utilisée dans les circuits fonctionnant à très basse tension à cause de deux principaux désavantages :

- La faible robustesse aux variations environnementale comparé aux structures maitre-esclave - problème traité au chapitre 3,
- le temps de maintien positif qui induit une surconsommation énergétique de par l'insertion de tampons en délais - problèmes traités au chapitre 4.

Au chapitre 2, une comparaison juste et rigoureux de six architectures, semblant les plus prometteuses de l'état de l'art, a été effectuée dans le domaine énergie-délais pour mettre en évidence l'architecture la plus efficace en énergie. Si l'architecture TGPL-Clk présente la meilleure efficacité énergétique pour les opérations haute vitesse, c'est l'architecture C$^2$MOS-Data qui s'est révélée l'architecture de bascule à impulsion la plus efficace en énergie sur une large gamme de délais et tension d'alimentation.

Dans le chapitre 3, le compromis fondamental entre la robustesse et la consommation énergétique d'une bascule à impulsion est expliqué et un générateur de délais (GD) utilisant la technique du *current-starved* a été proposé pour surmonter ce problème. Il a été montré que notre GD proposé améliore sensiblement la robustesse des bascules à impulsion, ce qui était l'un des deux désavantages majeurs. De plus, cette structure est très flexible et offre plusieurs degrés de liberté aux concepteurs de circuit. Des mesures sur silicium ont permis de comparer le rendement fonctionnel et la tension d'alimentation ($V_{dd}$) minimum de différentes bascules, et ont montré que notre architecture de GD permet d'atteindre une augmentation du rendement moyen de 7.5% pour une mme $V_{dd}$ et un rendement moyen identique obtenu à une tension d'alimentation jusqu'à 45mV plus petite que sans la technique du *current-starved*.

Dans le chapitre 4, une technique de capture conditionnelle a été présentée et implémentée dans la bascule efficace en énergie du chapitre 2. Ensuite, une comparaison dans le domaine énergie-délais a montré que cette architecture de bascule à impulsion présente une dissipation d'énergie plus petite que les topologies maitre-esclaves, et un temps de maintien positif. Le second inconvénient des bascules à impulsion est donc partiellement résolu grce à cette architecture. Finalement, un banc de registre, basé sur les bascules robustes et efficaces en énergie provenant des innovations des chapitres précédents, est comparé à un banc de registre contenant des bascules maitre-esclaves. Cette comparaison a montré que nos bascules à impulsion explicite fournissait une surface et une consommation d'énergie inférieures, tout en garantissant une robustesse suffisante dans le régime sous-seuil.

En conclusion, nous avons conçu des architectures de bascules à impulsion explicites très efficace en énergie, à savoir la TGPL-Clk, C$^2$MOS-Data et C$^2$MOS-Xor, dédiées respectivement à la très haute performance, l'efficacité énergétique et la très faible puissance. La technologie FDSOI, à travers la technique de polarisation face arrière, permet de modifier dynamiquement les performances en

énergie et en délais des bascules, en fonction des contraintes courantes du circuit. L'efficacité énergétique est préservée durant la synthèse grce à la technique de capture conditionnelle présentée. Finalement, la robustesse à très basse tension est assurée par la faible variabilité du FDSOI, le choix adéquat de la polarisation face arrière - grce à notre étude du rendement fonctionnel - et notre architecture innovante de générateur de délais.