Instituto Tecnológico de Costa Rica

Escuela de Ingeniería Electrónica

VOLTAGE DROP TOLERANCE BY ADAPTIVE VOLTAGE SCALING USING CLOCK-DATA COMPENSATION

Documento de tesis para optar por el grado académico de Maestría en Electrónica con énfasis en VLSI

Andrés Malavasi Mora

8 de mayo de 2019

Instituto Tecnológico de Costa Rica

Escuela de Ingeniería Electrónica

Tesis de Maestría

Tribunal evaluador

Tesis de maestría defendida ante el presente Tribunal Evaluador como requisito para optar por el grado académico de maestría, del Instituto Tecnológico de Costa Rica.
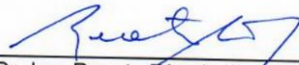
Miembros del Tribunal

Dr. Roberto Pereira Arroyo

Profesor lector

Máster Ronny García Ramírez

Profesor lector

Dr.-Ing. Renato Rímolo Donadío

Director de Tesis

Los miembros de este Tribunal dan fe de que la presente tesis ha sido aprobada y cumple con las normas establecidas por la Escuela de Ingeniería Electrónica.

Cartago, 07 de mayo de 2019

I declare that this thesis document has been made entirely by myself, using and applying literature related to the subject and introducing knowledge and experimental results of my own.

In the cases in which I have used bibliography, I have proceeded to indicate the sources through the respective bibliographic citations. Consequently, I assume full responsibility for the thesis work carried out and for the content of this document.

Andrés Malavasi Mora

Cartago, October 7th, 2013

ID: 1-1343-0344

Resumen

El ruido de alta frecuencia en la red de alimentación compromete el rendimiento y la eficiencia energética de los sistemas electrónicos con microprocesadores, restringiendo la frecuencia máxima de operación de los sistemas y disminuyendo la confiabilidad de los dispositivos. La frecuencia máxima será determinada por la ruta de datos más crítica (la ruta de datos más lenta). De esta manera, es necesario configurar una banda de guarda para tolerar caídas de voltaje sin tener ningún problema de ejecución, pero sacrificando el rendimiento eléctrico.

Este trabajo evalúa el impacto de la caída de voltaje en el rendimiento de los circuitos CMOS de alta densidad, estableciendo un conjunto de casos de prueba que contienen diferentes configuraciones de circuitos. Se desarrolló una técnica adaptable y escalable para mejorar la tolerancia a la caída de voltaje en los circuitos CMOS a través del escalado adaptativo, aprovechando el efecto de compensación de datos del reloj. La solución propuesta se validó aplicándola a diferentes casos de prueba en una tecnología FinFet-CMOS a nivel de simulación del diseño físico.

Palabras clave: caída de voltaje, escalamiento adaptativo de voltaje, compensación de datos-reloj, ruido en la red de distribución de alimentación

Abstract

High-frequency power supply noise compromises performance and energy efficiency of microprocessor-based products, restricting the maximum frequency of operation for electronic systems and decreasing device reliability. The maximum frequency is going to be determine by the most critical data path (the slowest data path). In this way, a guard band needs to be set in order to tolerate voltage drops without having any execution problem, but leading to a performance reduction.

This work evaluates the impact of voltage drop in the performance of CMOS circuits by establishing a set of test cases containing different circuit configurations. An adaptive and scalable technique is proposed to enhance voltage drop tolerance in CMOS circuits through adaptive scaling, taking advantage of the clock-data compensation effect. The proposed solution is validated by applying it to different test cases in a FinFet CMOS technology at a post-layout simulation level.

Keywords: Adaptive voltage scaling, clock-data compensation, voltage drop, power noise.

Dedico este trabajo a Dios y a mi familia por sus constantes bendiciones en la vida, por guiarme a lo largo de este camino, ser el apoyo y fortaleza en aquellos momentos de dificultad y de debilidad.

Acknowledgments

First, I would like to thank my thesis tutor Renato Rimolo Donadio, for showing interest and support on this project and giving me the opportunity to work with him throughout this investigation. I would also like to thank all the support received from the administrative staff of the Electronics Master Program of Costa Rica Institute of Technology, especially Grettel Trejos Salas and Anibal Coto Cortés, who were always attentive and willing to help during the duration of the master's degree program.

Also, I acknowledge all the support provided by my co-workers at Circuit Research Lab-Intel Labs, especially to Carlos Tokunaga and James Tschanz for their constant teachings and guidance throughout this work, which helped to significantly improve the quality of this research and guide it in the right direction.

Finally, I would also like to acknowledge all that effort and support from my family and friends, who always showed interest and support during this thesis process, always trusting in my abilities and believing that this moment was going to come.

Andrés Malavasi Mora

Cartago, 8 de mayo de 2019

Contents

List of Figures

List of Tables

Introduction

High-frequency supply voltage (V_{CC}) drop degrades the performance and energy efficiency of microprocessor products, limiting the maximum frequency (f_{max}) of operation for electronic systems such as microprocessors [1]. This f_{max} is going to be determined by the most critical data path (the slowest data path). This is why, a guard band needs to be set in order to tolerate voltage drops without having any execution problem, but leading to a performance penalty.

In the past, several adaptive circuit techniques have been reported in the literature aiming to reduce the effect of the voltage drops by explicitly sensing the variation with on-die monitors and adjusting the operating condition (e.g., clock frequency). Although this is effective at low frequencies, the chances to mitigate high frequency drop are very limited. Other techniques like resilient timing-error detection and recovery circuits can be very useful since they detect the timing violation, isolates the error from corrupting the architecture state, and corrects the error through instruction replay [2]–[5]. A resilient design is highly effective at mitigating the impact of high-frequency drops on performance. However, the architectural design complexity for implementing error recovery into a high-performance microprocessor while ensuring coverage for all failure scenarios is a significant challenge.

Adaptive clock distribution (ACD) along with clock gating techniques have become an interesting solution to mitigate the effect of voltage drop on microprocessor performance [1]. These techniques take advantage of the clock-data compensation effect, in which both clock and data are affected by the drop: changing the clock signal compensates changes on data paths. Nevertheless, clock gating can bring synchronization problems between blocks in a high performance microprocessor.

This work evaluates the impact of voltage drop in the performance of CMOS circuits. An adaptive and scalable technique is developed to enhance voltage drop tolerance in CMOS circuits through adaptive voltage scaling, taking advantage of the clock-data compensation effect. First, a series of different circuit configurations are analyzed to understand the impact of this phenomenon. The proposed solution is validated by applying it to different scenarios, considering process corners and operating conditions in FinFet-CMOS technologies. Finally, layout solutions and their evaluation were incorporated in order to have higher confidence on the results with the proposed technique: up to 22.4% frequency improvement and 32.5% tolerance for extra current consumption of the circuit; at the same time, frequency operation up to 770 MHz was stablished for the studied configurations as the upper bound, which implies that the load at the frequency must be restricted to be equivalently equal or slower than this bound.

The document is organized as follows: Chapter II explains the impact of high-frequency voltage drop in high performance microprocessors and also describes in detail the clock-data compensation effect based in the available literature. Chapter III defines the initial clock-data compensation studies, which set the baseline comparison with the proposed circuits. Chapter IV presents the proposed technique to control the voltage when a voltage drop is happening in the system. Finally, Chapter V addresses layout implementations for the circuits designed on Chapter III and Chapter IV, followed by parasitic extraction and post-layout simulations, creating a model of the system closer to reality with results that are more accurate.

Chapter II

Power Supply Noise Fundamentals

Power supply noise in the form of voltage variations arises due to *IR* drop and *Ldi/dt* events. The *IR* drop has been increasing over time due to increased resistances in the power grid as the metal widths continue to shrink with each successive technology generation. The inductive *Ldi/dt* events are also increasing due to the higher current demands in more complex chips. When it comes to overcome voltage supply noise effects due to *Ldi/dt* events, it is good to understand all the aspects involved in the power distribution network and their consequences.

2.1. Impact of Power Noise

Ground bounce and V_{CC} bounce, also known as power noise, have always been present in digital circuits. However, in the past they were not always noticeable because of the complexity of the circuits, having very low operating frequency, small pin count and simple functionality [38]. However, modern processes are becoming more sensitive to noise. In addition to technology parameters having a larger variation with each new technology generation, timing sensitivity to such environmental conditions as temperature, aging, workload, cross-talk noise in wires, and many other effects, is increasing.

Noise processes that affect timing (performance) are described as random or systematic. Random noise is less dependent on the integrated circuit's design than systematic noise and it is characterized by a number of statistics such as its mean and standard deviation under the assumption of a normal distribution. Systematic noise results from characteristics of the manufacturing process or from the physical design and can be estimated once the underling process causes the variation is understood. Once the source of systematic variation is identified, designs can be adjusted or processing can be modified to reduce variation [39].

As the operating frequency increases, the average on-chip current required to charge (and discharge) these capacitances also increases, while the time switching current interval decreases. Therefore, a large change in the total on-chip current occurs within a short period of time. The primary sources of the current surges are the input/output (I/O) drivers and the internal logic circuitry, particularly those gates that switch close in time to the clock edges. Because of the self-inductance of the off-chip bonding wires and the on-chip parasitic inductance inherent to the power supply rails, the fast current surges result in voltage fluctuations in the power supply network, which is called simultaneous switching noise (SSN) or delta-I noise (*di/dt*) [30].

The SSN voltage is related, as a first order approximation, to the inductance present between the device ground and the system ground, and the current across this inductance. The most important parameter to estimate is the peak noise, which can be derived using the standard expression for the voltage across an inductor, expressed as:

$$\Delta V = I_{supply} * R_{mesh} + L_{pack} \frac{dI_{supply}}{dt} \tag{2.1}$$

where R_{mesh} is the power grid (mesh) resistance, L_{pack} is the package and pin inductance, and I_{supply} is the current flow through the user logic circuits and ΔV is the noise appearing on device ground relative to system ground in Figure 2.1:
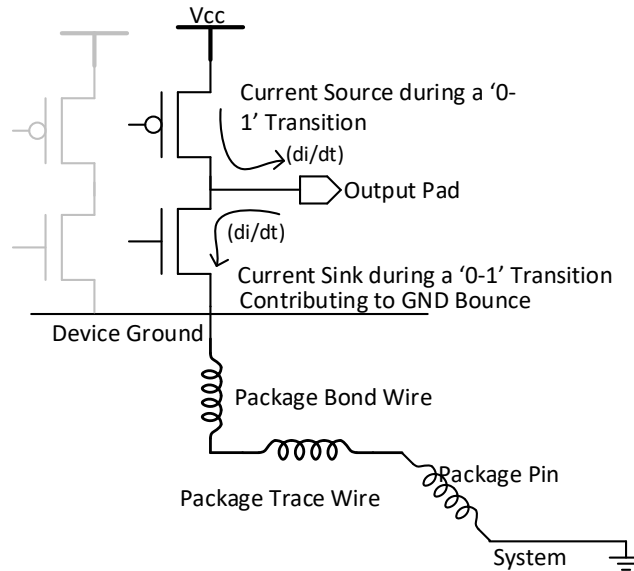
Figure 2.1: A Sample Switching Output Buffer Showing Parasitic Inductance

Therefore, the higher *di/dt*, the higher the ground bounce amplitude. The device ground is connected to the system ground (PCB ground) through a series of inductors, comprised of package bond wire, package trace, and board inductance. As a result, the higher L_{eff}, the higher the amplitude will be.

$$L_{eff} = L_{bondwire} + L_{trace} + L_{pin} \tag{2.2}$$

Initial research work was oriented towards estimating L, using a simple model of *di/dt* or assuming that *di/dt* was extracted from simulation. One simple model of *di/dt* is the triangle approximation. It is assumed that the current has a triangular shape as in Figure 2.2, with peak I_p and duration t_f, the fall time of the discharged node. T to denote the amount of time required for the current to reach its peak value. *di/dt* noise levels are smallest when $T = t_f/2$, since this value minimizes the steepness of both the rising and falling edges.

$$v_{nmax} = L * \left(\frac{di(t)}{dt}\right)_{max} = L * \frac{I_p}{T} \tag{2.3}$$

Equation 2.3 describes the peak noise that would show up on Vss, depicted in Figure 2.2 when the output node, is switched from high to low. Due to symmetry, the same noise would show up on

Vcc when the output buffer is switched from low to high. When internal circuits are switching, Vss and Vdd will simultaneously have symmetric noise (assuming equal inductance in the Vcc and Vss supply) such that we can have equation 2.4 and 2.5 to represent the device supplies levels:

$$Vcc_{min} = pwr - v_{nmax} \qquad (2.4)$$
$$Vss_{max} = gnd + v_{nmax} \qquad (2.5)$$

where pwr is the system power supply and gnd is the system ground.
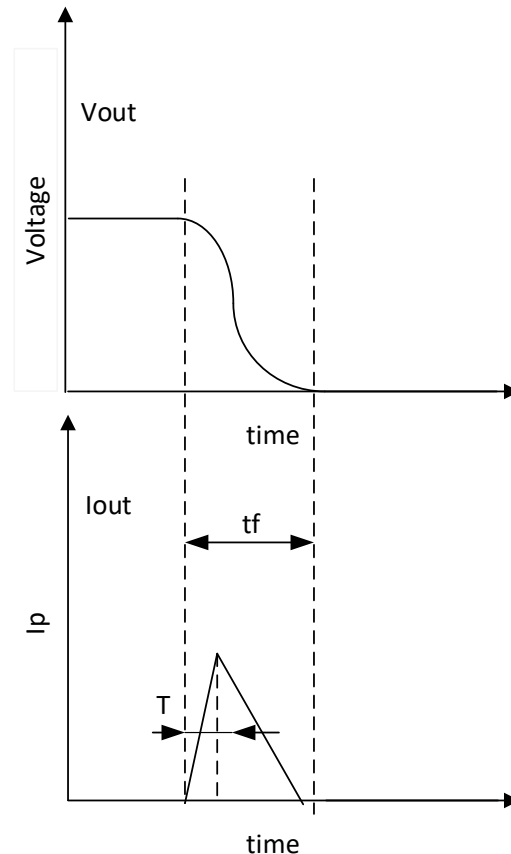
Figure 2.2: Triangle approximation of current

Problems may arise when this ground bounce gets transferred to the outside through output buffers driving low. If the bounce is high enough, there is a possibility that the glitch will be recognized as a legal logic '1'. The same phenomenon applies for Vcc bounce when driven a logic "1", but devices usually tend to have more noise margin near the high level ('1') than near the low level ('0'). Therefore, ground bounce is considered more often [38].

Simultaneous switching noise (SSN) has become an important issue in the design of the internal on-chip power distribution networks in current very large scale integration/ultra large scale integration (VLSI/ULSI) circuits [30]. Transition from bipolar to CMOS processing led the increase of di/dt noise levels in integrated circuits. As CMOS technology is scaled to give smaller and faster transistors, the power supply voltage must decrease. Having high performance processors requires higher operating frequency, hence clock rates rise and the power consumption must increase. Any designer working with high-edge-rate devices must be aware of these noise

issues and will need to address them. With increasing clock speeds and decreasing supply voltages in computing devices, the correct design and measurement verification of power-distribution networks (PDN) become more challenging [32].

Design of the power distribution system (PDS) is becoming an increasingly difficult challenge for modern CMOS technology. As CMOS technology is scaled to give smaller and faster transistors, the power supply voltage must decrease. As clock rates rise and more function is integrated into microprocessors and application specific integrated circuits (ASIC's), the power consumed must increase [40]. The power distribution subsystem of a chip consist of metal wires or planes on the chip, in the package, and the printed circuit board. It also includes bypass capacitors to supply the instantaneous current requirements of the system. An ideal power distribution network has the following properties:

- Maintains a stable voltage with little noise
- Provides average and peak power demands
- Provides current return paths for the signals
- Avoids wear out from electromigration and self-heating
- Consumes little chip area and wiring
- Easy to lay out

Real networks must balance these competing demands, meeting targets of noise and reliability as inexpensively as possible.

The on-chip power distribution network consists of power and ground wires within the cells and more wires connecting the cells together. Most cells contain internal power and ground busses routed on the available metal layers. These wires are typically wider than minimum to provide lower resistance and better electromigration immunity, being upper layers with wider metals used to route the higher level grids. The power grid extends across the entire chip or voltage domain. Ultimately, it must connect to the package through the I/O pads. When a pad ring is used, the connections are all near the periphery of the chip.

The resistance of the power supply network includes the resistance of the on-chip wires and vias, the resistance of the bond wires or solder bumps to the package, the resistance of the package planes or traces, and the resistance of the printed circuit board planes. Because the package and printed circuit board typically use copper that is much thicker and wider than on-chip wires, the on-chip network dominates the resistive drop (IR drop). IR drops arise from both average and instantaneous current requirements. The instantaneous current may be much larger than the average drop because current draw tends to locally spike near the clock edge when many registers and gates switch simultaneously.

Chips need a substantial amount of capacitance between power and ground to provide the instantaneous current demands of the chip. This is called *bypass* or *decoupling capacitance*. The bypass capacitance is distributed across the chip so that a local spike in current can be supplied from nearby bypass capacitance rather than through the resistance of the overall power grid. It also greatly reduces the di/dt drawn from the package. Bypass capacitance near the switching gates can supply much of this instantaneous current, so a well-bypassed power supply network only

needs low enough resistance to deliver the average current demand, not necessarily the peak. The only dielectric available in a standard CMOS process to build compact high-capacitance structures is gate oxide, so the extra bypass capacitance is commonly built with an nMOS transistor with the gate tied to Vcc and the source and drain tied to GND. Decoupling capacitor layout should maximize the capacitance per unit area.

The power supply in high complexity CMOS circuits should provide sufficient current to support the average and peak power demand within all parts of an integrated circuit. An inductive, capacitive, and resistive model is used in this section to characterize the power supply rails when a transient current is generated by simultaneous switching of the on-chip registers and logic gates within a synchronous CMOS circuit. Figure 2.3 shows a lumped model of the power distribution network for a system including the voltage regulator, the printed circuit board planes, the package, and the chip. The network also includes bypass capacitors near the voltage regulator, near the chip package, possibly inside the chip package, and definitely on chip. The external capacitors are modeled as an ideal capacitor with an effective series resistance (ESR) and effective series inductance (ESL) representing the parasitics of the capacitor package. Larger capacitors have bigger effective series inductances.
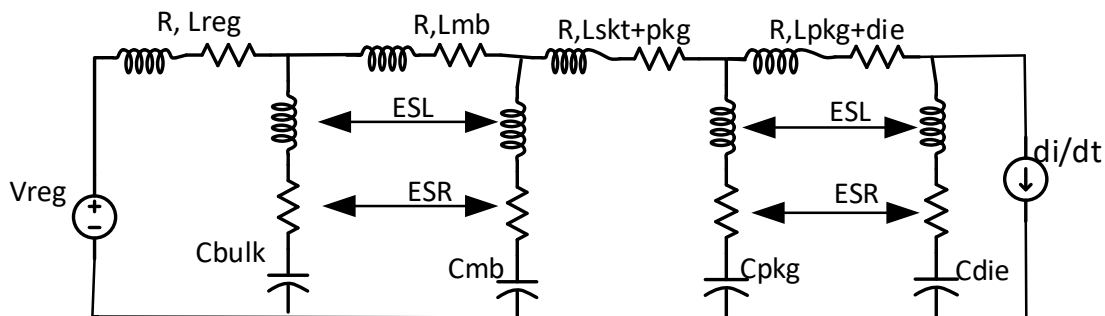
Figure 2.3: Power delivery system with multiple stages

The voltage regulator seeks to produce a constant output voltage independent of the load current. It is modeled as an ideal voltage source in series with a small resistance and the inductance of its pins. Near the regulator is a large bulk capacitor (typically electrolytic or tantalum). Power and ground planes on the printed circuit board carry the supply current to the package, contributing some resistance and inductance. Typically, the board designer places several small ceramic capacitors near the package. The package and its pins again contribute resistance and inductance. High-frequency packages often contain small capacitors inside the package for further decoupling. Finally, the chip connects to the package through solder bumps or bond wires with additional resistance and inductance.

The on-chip bypass capacitance consists of the symbiotic capacitance and possibly some explicit decoupling capacitance. It typically has negligible inductance because it is located close to the switching loads. Decoupling capacitors in each stage serve as local storage to supply charge to the next stage when quickly needed. From the regulator to the die, with progressively higher quality (e.g., smaller ESL, ESR) but lower valued decoupling capacitors, the coverage frequency increases. [9]

As one moves from the chip toward the voltage regulator, each capacitor typically increases by about an order of magnitude. However, each series inductance increases by a similar amount. [41] illustrates a representative power delivery network for a high performance 90 nm microprocessor. The capacitance is on the order of 1 µF on the die, 10's of µF in the package, and 100's of µF on the board, and 1 mF at the voltage regulator. The inductance is about 1 pH between the die and package, 10 pH between the package and board, and 100 pH along the board to the voltage regulator. The resistance is a fraction of an mΩ at each link.

Power delivery system of a microprocessor ideally strives to maintain a low constant impedance across all frequencies. In practice, this necessitates several stages of decoupling to optimally flatten the supply impedance across a broad range of frequencies [9]. If the system draws P watts of power and the maximum allowable power supply ripple is $r \times V_{cc}$ (e.g., $r = 0.1$ for 10% supply noise), then the supply impedance must be less than

$$Z = r * \frac{V_{CC}^2}{P}$$
(2.6)

This relationship shows that required supply impedance is dropping quadratically with voltage scaling. It is also dropping as power consumption increases. This impedance requirement has driven the adoption of improved packages and flip-chip bonding with solder bumps instead of bond wires. It means chips need to use more metal and on-chip bypass capacitance. For example, a 1.0 V system dissipating 100 W of power draws 100 A. To keep supply noise down to 10% of Vcc, the power supply impedance must be 1 mΩ. If the system had no bypass capacitance, the distribution network would consist of only the resistance and inductance, so it would have an impedance of $Z = R + j\omega L$. This impedance increases with frequency ω and becomes unacceptably high for most systems by about 1 MHz.

The bypass capacitors in parallel with the supply provide an alternative low impedance path at higher frequencies. An ideal capacitor has impedance that decreases with frequency as $Z = 1/j\omega C$. Unfortunately, the effective series inductance of the capacitors limits the useful frequency range of the real capacitor. The impedance of a capacitor C with effective series resistance R and inductance L is

$$Z = R + j\omega L + \frac{1}{j\omega C}$$
(2.7)

This impedance has a minimum of $Z = R$ at the self-resonant frequency of

$$f_{resonant} = \frac{\omega_{resonant}}{2\pi} = \frac{1}{2\pi\sqrt{LC}}$$
(2.8)

Figure 2.4 plots the magnitude of the impedance of a 1pF capacitor with 0.25nH of series inductance and 0.03Ω of series resistance. The capacitor has low impedance near its resonant frequency of 10 MHz, but higher impedance elsewhere.
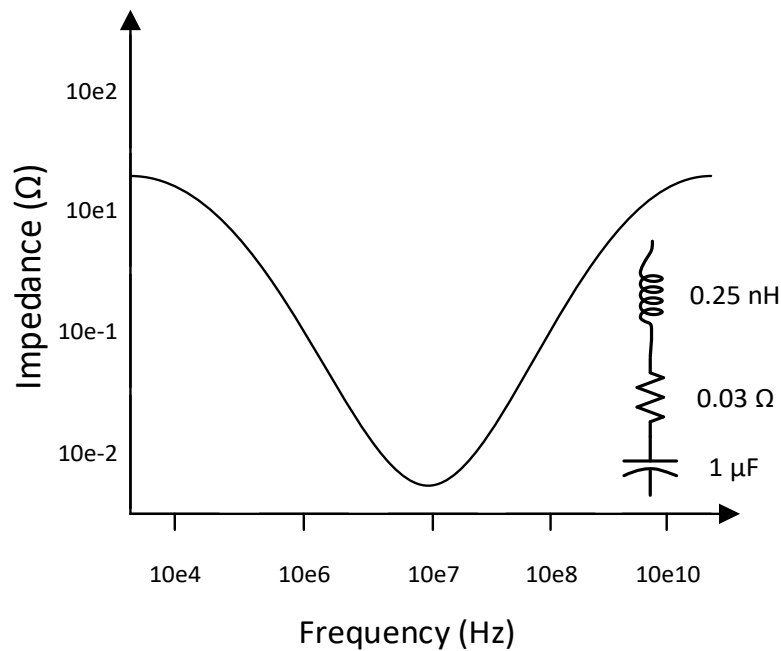
Figure 2.4: Impedance of a 1 uF bypass capacitor

Larger capacitors tend to have higher effective series inductances and therefore have lower self-resonant frequencies beyond which they are not useful. Thus, the system uses many capacitors of different sizes to provide low impedance over all the frequencies of interest.

Today, the improvements in packaging technology are barely sufficient to keep noise to tolerable levels. Even with exotic packages tailored for low noise, it is necessary to estimate noise levels and design for low noise before starting production of an integrated circuit [31]. Also, for the die supply it is impractical to near perfect filtering placing die capacitance, this will increase area and cost of a microprocessor.

V_{CC} drops primarily affect circuits globally across the die and may occur with frequencies ranging in delay from a few nanoseconds (i.e., high frequency) to a few microseconds (i.e., low frequency) [1]. Power supply noise can be summarized in 3 different types of drops, each of them has different magnitude and frequency components. Figure 2.5 shows the simulated voltage drops on a typical FC-PGA Pentium 4 processor (130 nm) system from an anticipated architectural event. [9]
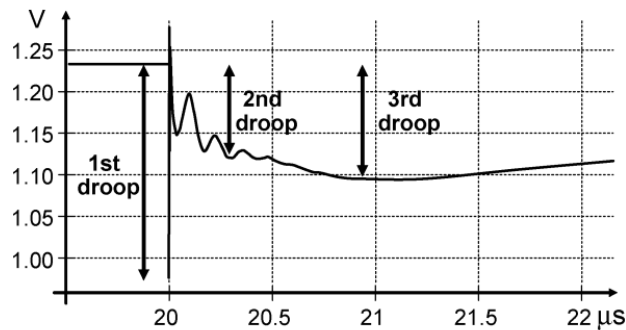
Figure 2.5: Simulated voltage drops on a typical FC-PGA Pentium 4 processor (130 nm) system from an anticipated architectural event. *From [9]*

The most critical drop is the first one, it is caused by the *LC* tank formed between the package/bonding inductance and the die capacitance and thus affects the entire chip and its duration is of a few nanoseconds. When the current demand suddenly increases, the extra charge is initially drawn out of the on-chip bypass capacitors. As these capacitors discharge, the supply voltage drops precipitously. Although this is the shortest drop, it is also usually the deepest and therefore can severely impact microprocessor performance when a critical path is accessed in conjunction with the drop. The delay depends on the inductance of the bumps. Moreover, this inductance may cause the supply voltage to overshoot and oscillate. Meanwhile, the capacitors in the package supplying this current start to discharge and the voltage drops again. This second drop occurs on a longer time scale determined by the package capacitance. Eventually, the current through the package pins and socket increases to begin recharging the package capacitors. Meanwhile, the capacitors on the printed circuit board discharge, leading to a third drop before the voltage regulator catches up with the increased current demand. The second and third drops are minimized by providing an adequate number of high-quality, low ESL capacitors at each stage in the power distribution network [40].

Designers typically assume that adding on-chip bypass capacitance to reduce supply drop improves operating frequency. While more capacitance certainly does reduce the drop, the frequency does not necessarily improve. In a striking experiment, Wong et al. in previous publications fabricated several wafers of Pentium 4 processors with and without decoupling capacitors. Without capacitors, the first drop increased by 8% of Vcc, but the operating frequency only slowed by 1%. The anomaly was explained by showing that under certain conditions the noise modulates the clock period in a way that tracks the critical path delay [9].

The first drop (also called "first drop noise") can be excited by a sudden current spike caused by a clock edge or a processor wakeup operation [10]. Previous publications [11]–[13] have showed this supply noise is between 50 MHz to 300 MHz resonant frequency band, being the dominant noise component in a typical high performance microprocessor.

The first drop due to its frequency components, it is consider as high frequency supply noise. Figure 2.6 shows the measured supply network impedance of an Intel Nehalem microprocessor, which exhibits a large impedance peak at around 150 MHz [14].
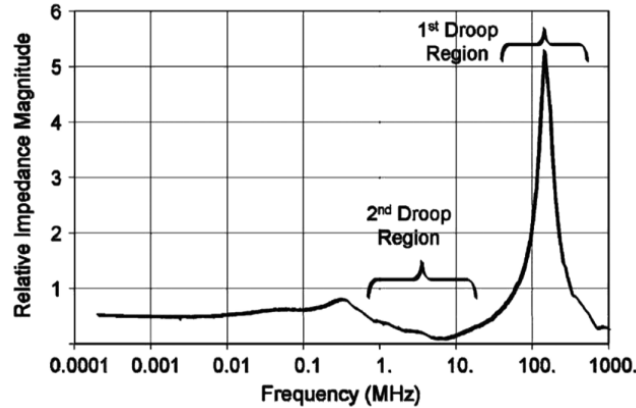
Figure 2.6: Measured supply network impedance response of Intel's Nehalem. *From [14]*

High Frequency supply voltage (V_{CC}) drop degrades the performance and energy efficiency of microprocessor products, limiting the maximum frequency (fmax) of operation for electronic systems such as microprocessors. This fmax is going to be determine by the most critical data path (the slowest data path). This is why a guardband needs to be set in order to tolerate voltage drops without having any execution problem, but leading to a performance reduction.

2.2. Impact of High-Frequency Voltage Drop

Power-supply noise on the performance of high-frequency microprocessors can reduce oxide reliability [6] and make SRAM cells unstable [7] besides of impacting performance. Most of the literature concerned with the timing impact of the noise concentrates on the maximum voltage fluctuation created by the simultaneous switching of a large number of transistors. From a system perspective however, some authors argue that the worst-case power-supply noise is only indirectly important [8]. The reason is that in practice, the power-supply noise tends to impact timing before causing other failures. When this is the case, the performance impact of the noise (in GHz) is more relevant than its magnitude (in mV). For typical circuits, it is shown that the peak of the noise is largely irrelevant and that the average supply voltage during switching is more important. It is then argued that global differential noise can potentially have a greater timing impact than common-mode noise. Circuit arguments are then given to justify why the global interconnects used for clock distribution are particularly critical.

There are several scenarios where voltage drops can be harmful and cause errors on integrated circuits:

A. PLL jitter
A phase-locked loop (PLL) is incorporated into many integrated circuits for skew reduction and frequency multiplication. The core of a PLL consists of a VCO with finite power supply rejection (PSR), i.e. the frequency of the VCO is somewhat dependent on V_{CC}. A change of VCO frequency due to V_{CC} noise can give a critical path error. One clock edge generated by the VCO triggers the start point flip-flop (generating flip-flop) of a timing path. If V_{CC} noise increases the VCO

frequency, the next clock edge will arrive at the endpoint flip-flop (capturing flip-flop) before the logic has completed evaluation, and thus causing a logic error. Peak cycle-to-cycle jitter is a PLL metric that quantifies the robustness against timing errors.

If the power supply noise is extended over several periods of the VCO clock, jitter accumulation will occur such that each clock edge will deviate more and more from the ideal edge location. This may not lead to a critical path problem, but can give synchronization failures between different clock domains. A more relevant PLL metric in this case is peak-to-peak jitter.

B. I/O reference levels

Input buffers have a trigger level, such that the input is interpreted as a "1" when the input voltage is higher than the trigger level and otherwise "0". Power supply noise coupled through an input buffer affects the trigger level, so that the input can be misinterpreted to an incorrect logic level [34]. Even if the noise does not cause misinterpretation, the reduced voltage between the input signal and the trigger level causes speed degradation, which may give timing failures. Similar problems may arise when the output buffer of a transmitting chip experiences power supply noise.

C. Dynamic logic

Dynamic logic is characterized by having nodes that occasionally are in a high-impedance state storing logic values as charge on capacitors. If the gate-to-source voltage on a high-impedance transistor experiences noise, a dynamic node might be erroneously charging/discharging [34].

D. Logic delay failure

This investigation is focused on this kind of scenario, in which a critical timing path of a pipelined circuit is shown in Figure 2.7, being affected by a voltage drop.
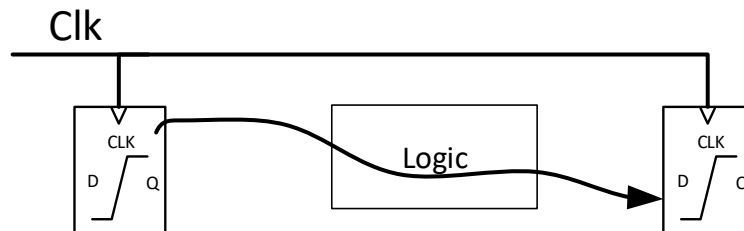
Figure 2.7: Voltage drop may cause delay errors in pipelined circuits.

The delay in the path is proportional to the power and ground levels. If there is a bounce on V_{SS}, and/or a negative glitch on V_{CC}, the critical path might not complete evaluation before the receiving flip-flop is triggered by a clock edge, causing either setup or hold violations.

Present delay models that take this noise into account are proposed for two different kinds of paths: device-dominated and interconnect-dominated timing paths. Usually device-dominated paths shares same power and ground voltages, causing if there is any voltage drop, it is going to affect all the gates involved in the timing path. On the other hand, interconnect-dominated paths are almost always distributed since they are constructed using long wires. Hence, gates involved in

the path can have different power domain, and if there is any voltage drop present, probably its effect it won't be the same for all the gates.

Interconnect-dominated paths are constructed by cascading repeaters and long wires. These repeaters are typically inverters or buffers. The length, width, and spacing of the wires together with the transistor sizes of the repeaters are optimized during design. The goal is usually to maximize performance under metal usage, power, and noise constraints. This optimization process usually makes the repeater and wire delays comparable [20]. Interconnect-dominated paths tend to be less sensitive than device-dominated paths to chip-scale power-supply noise events. The noise directly affects the delay and the output edge rate of the repeaters. However, when it modifies the output edge rate of the repeaters, the noise only has an indirect effect on the delay of the wires. The problem is that some interconnect-dominated paths, in particular the ones used for clock distribution, are getting increasingly long. Since clock frequencies are also increasing, the number of clock cycles spanned by these paths is increasing even faster. This is a problem because the number of clock cycles actually multiplies the timing impact of the supply voltage variations.

For instance, say that a given voltage variation causes a delay increase of 6% for a one-cycle device-dominated path and a 4% increase for a two-cycle interconnect-dominated clock path. Then, the delay of the clock path varies by of one clock cycle. Thus, for long interconnect-dominated paths, a given amount of noise can cause a delay variation that is a large fraction of the clock cycle, despite being small with respect to the path delay. When the delay of a clock path varies, clock inaccuracy (i.e. skew or jitter) is created. In a data path through a network of sequential elements, the interval between the data launch and its final capture determines the time available for performing computations. If the clock of the launching sequential element is late or if the clock of the capturing element is early, the timing margin of the path is directly reduced. If the margin becomes negative, the frequency must also be reduced. Consequently, the performance impact of the power-supply noise on the clock distribution network is highest when it creates a combination of slow and fast (i.e. unbalanced) clock paths.

For instance, in the Pentium 4 microprocessor, all clock paths are chip-wide. When the supply voltage is not uniform, two paths routed over different regions of the die will experience different voltages and therefore exhibit different delays. This makes the clock distribution network particularly sensitive to global differential power-supply noise. If the clock paths become long enough, the differential power-supply noise can create enough clock inaccuracy to make its frequency impact greater than the one of the common-mode noise on device-dominated paths [44].
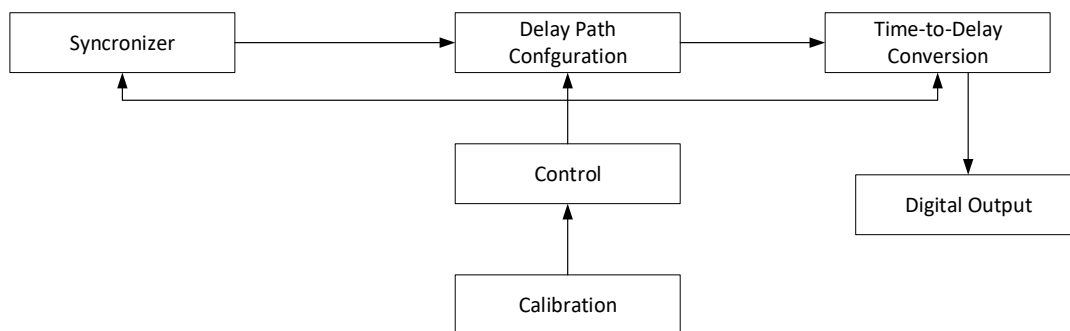
Figure 2.8: A simplified block diagram showing the basic building blocks inherent in most published critical path monitors

The operation of a critical path monitor like in Figure 2.8 is as follows: the system clock triggers the launch of a timing signal into a delay path; after the delay of the clock period, the phase of the timing signal and the system clock is captured by some time-to-digital conversion and compared to the expected phase; the difference between the captured and the expected phase indicates the amount of slack available in the timing. A block of logic is added to control the critical path monitor for operation and testing, and calibration data is maintained to provide the needed sensor accuracy.

2.3. Increasing Voltage Tolerance by Modifying Passive Devices

The PDN is responsible for distributing the power and ground voltages to all devices of the design. In previous sections, it was discussed that these components have parasitic inductances, resistances and capacitances. It good to understand what the role of these parasitics is in the PDN, and how can we modified these in order to increase the robustness of our system. Also, it is useful to understand if modifying these components will help to overcome the high frequency voltage drops.

2.3.1. Reducing Inductance

Based on the information we presented above, there are several options to make our PDN more robust to high frequency voltage drops. Clearly, one option that can be explored is to reduce L_{pack}, but today, the improvements in packaging technology are barely sufficient to keep noise to tolerable levels. Even with exotic packages tailored for low noise, it is necessary to estimate noise levels and design for low noise before starting production of an integrated circuit [31].

2.3.2. Reducing resistance

Also, another option is to look reducing the resistance in the power grid by adding more power straps. However, this has two implications, the first implication is the overhead that this brings in terms of routing tracks available for the circuit. If our circuit is in a very congested area, routing tracks need to be share between power and signals, and there is a big possibility we do not have enough routing tracks to reduce the resistance significantly, the only way to do this is increase the area needed for the circuit, which incurs in an extra cost of the overall system.

Assuming designer do not care about the area and cost of the chip, in that case we will have enough routing tracks available to put more power straps, reducing R_{mesh} definitely will reduce ΔV, however, this needs to be reduced to levels in which the voltage drop will not cause any failure in the circuit. Now, this do not change much the frequency response during a di/dt event, the resistance will only change the DC response of the voltage supply (see Figure 2.9).
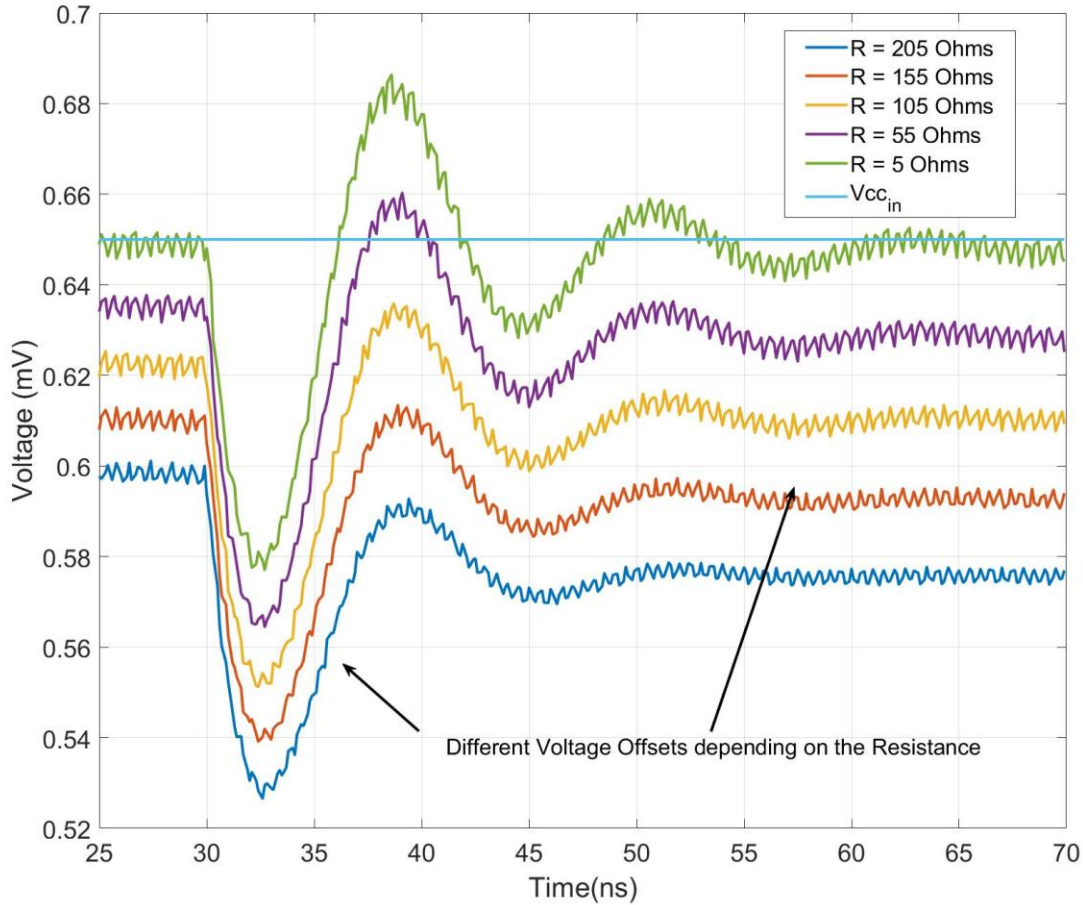
Figure 2.9: Resistance effect on voltage Supply under voltage drop

Therefore, the voltage drop still has the high frequency component, only with a reduced voltage drop. Moreover, if we have a bigger di/dt event, voltage drop will still make the circuit to fail.

2.3.3. Increasing capacitance

Finally, die capacitance can be increased. This die capacitance is driven mainly by active and passive devices as we mentioned before. Increasing the load capacitance is not an option since this is given by the circuits we are designing, in order words the only option we have is to increase the passive capacitance. Also, we are assuming the passive capacitance will overcome the load capacitance, otherwise, there won't be room to do some improvement in C_{die}.

Having said that, designers typically assume that adding on-chip bypass capacitance to reduce supply drop improves operating frequency. While more capacitance certainly does reduce the drop, the frequency does not necessarily improve. As mentioned before, Wong et al. [9] in previous publications fabricated several wafers of Pentium 4 processors with and without decoupling

capacitors. Without capacitors, the first drop increased by 8% of Vcc, but the operating frequency only slowed by 1%.

Therefore, once again we are not modifying the frequency of the voltage drop, we are modifying its magnitude, in which under a bigger di/dt event, voltage drop will still make the circuit to fail under high frequency voltage drops. Also, adding decaps to overcome the voltage drop problem has other implications like power, area and frequency response of the added capacitance.

In 90 nm and 65 nm processes, a new design issue for decaps due to oxide thickness reduction is the thin-oxide gate tunneling current. The current is in the form of tunneling electrons or holes from substrate to gate or from gate to substrate through the gate oxide, depending on the voltage biasing conditions. Two forms of gate tunneling exist: Fowler—Nordheim (FN) tunneling and direct tunneling. For normal operations on short-channel devices, FN tunneling is negligible, and direct tunneling is dominant [9]. In the case of direct tunneling, the gate leakage current in PMOS is much less than in NMOS, and it has been shown experimentally that PMOS gate leakage is roughly three times smaller than NMOS gate leakage for same size transistors. Assuming a 90nm technology with 1.7nm oxide thickness and 1.0V power supply, the gate leakage current I_{leak} is shown in Figure 2.10.

(a)   (b)
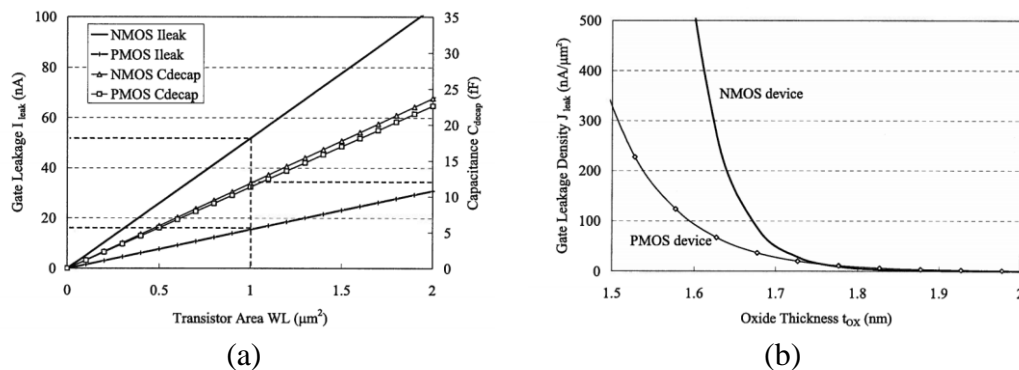
Figure 2.10: (a) Gate leakage current versus gate area; (b) Gate leakage current density J_{leak} versus oxide thickness t_{ox}

Clearly, as indicated from simulation results, the gate leakage is proportional to the transistor area. It is evident that at 90nm and 65nm technologies, the gate leakage from decaps will be significant [9]. The gate leakage contributes to the total static power consumption, and decaps usually occupy a large on-chip area. The use of PMOS devices exclusively is not a viable solution for high-frequency circuits since they have a poor frequency response relative to the NMOS devices for 90nm and 65nm.

In addition, the amount of gate leakage is also a strong function of the applied bias. If the transistor has a voltage across the oxide, V_{OX} set to close or below the threshold voltage V_{th}, it leaks significantly less. Indeed, under such a condition, the gate leakage current is typically 3-6 orders of magnitude less, depending on the values of V_{DD} and t_{OX}. Thus, the gate leakage in the second condition can be roughly considered to be zero. In decaps, the gate is at V_{DD} and the source and drain of a transistor are tied together. Therefore, decaps would experience the highest levels of leakage, as a function of V_{OX}.

Capacitance based alternative solution to mitigate voltage drop

As discussed before, any proposal related to an increase of the capacitance of the circuit will be too area and power costly. However, lets explore some of the capacitance based approach like the capacitance multiplier. Capacitor multipliers (CM) make low-frequency filter and long-duration timing circuits possible that would be impractical with actual capacitors. Another application is in DC power supplies where very low ripple voltage (under load) is of extreme importance.

In Figure 2.11a, shows the main concept of the CM, in which the current that is being draining from the source is M times i_C. In Figure 2.11b, the capacitance of capacitor C1 is multiplied by approximately the transistor's current gain (β). Without Q, R2 would be the load on the capacitor. With Q in place, the loading imposed upon C1 is simply the load current reduced by a factor of (β + 1). Consequently, C1 appears multiplied by a factor of (β + 1) when viewed by the load.

Now, since the BJT solutions is not applicable for the latest high-speed digital microprocessors, due to the technology compatibility, inFigure 2.11c, a CMOS based solution from [54]is depicted. In general, without going into details of the design of this circuit part of the drawbacks of the proposed solution is the use of extra capacitors that even if their value are going to be "multiplied", brings an extra overhead of area and power due to the reason we mentioned in the previous section. Second drawback is the proposed solution has a configuration in such way the drained current from the source is being multiplied by *M* times, functioning like a "bigger capacitance" than actually is; but the problem is this capacitor doesn't store the charge, basically redirects the charge from the source to the output node. So, once the voltage drop happens there is no way to deliver extra charge to circuit since the capacitor didn't store any further charge.

(a)

(b)

(c)
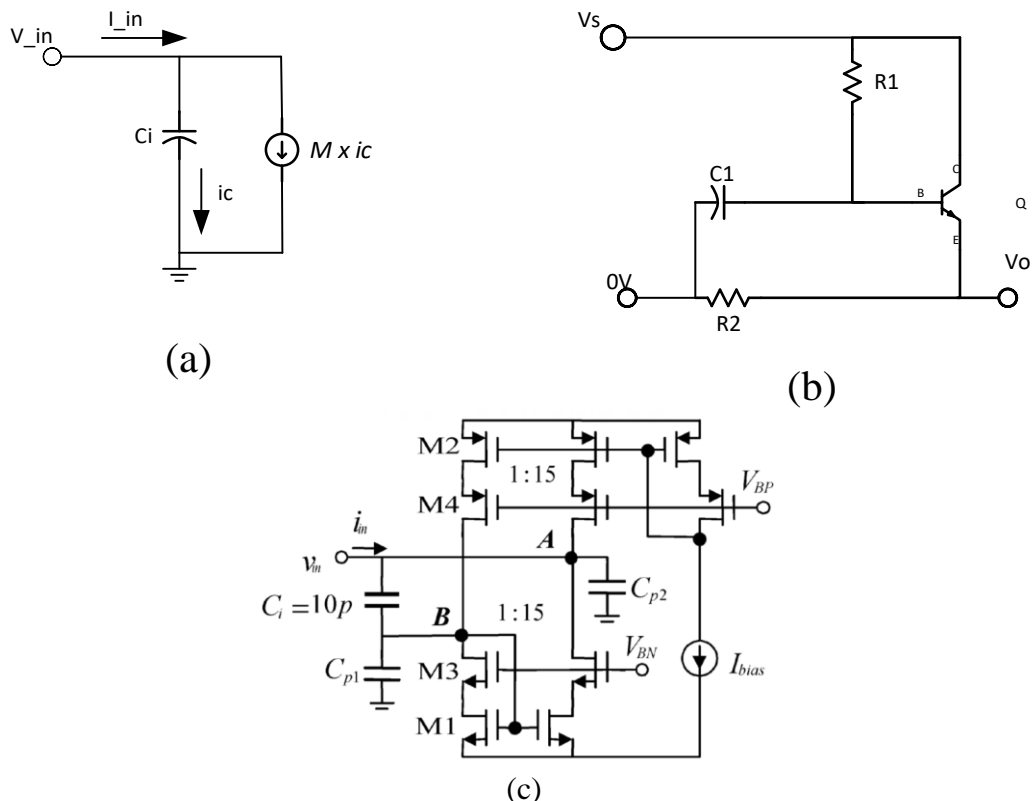
Figure 2.11: Capacitance multiplier: (a) Principle (b) BJT implementation (c) CMOS implementation (*From 54*)

2.4. Clock-Data Compensation Effect

Supply noise caused by on-chip current introduces delay variation in data paths, as well as jitter in clock paths. As a result, the launched data from one stage in a pipeline can no longer be guaranteed to be captured by the next clock edge within a given timing window (i.e., the clock cycle) leading to a timing failure [15].

Wong et al., in previous publications presented that there is beneficial interaction between the clock distribution and power supply noise with regards to microprocessor f_{max}. It is possible to enhance f_{max} immunity to power supply noise by designing the clock distribution such that the clock edge sampling the data at the receiver is pushed out every time the data signal is delayed due to power supply drop. They refer to this phenomenon as "Clock-Data Compensation" and denote it as CDC. [9]

This phenomenon is illustrated in Figure 2.12. A simple pipeline circuit consisting of a phase-locked loop (PLL), a clock path and a data path is shown. In traditional analysis, the clock period is assumed to be constant while only the data path delay is assumed to change under the influence of supply noise. Figure 2.12(b) illustrates example waveforms based on the traditional analysis showing several sampling failures during the event of a supply voltage undershoot. In reality, however, the clock path delay is also modulated by the supply noise and therefore stretches the clock period during supply downswings. As a result, the clock path delay and data path delay compensate for each other which alleviates the timing margin. Figure 2.12(c) shows example waveforms for this scenario in which the output is always sampled correctly by the stretched clock cycle.
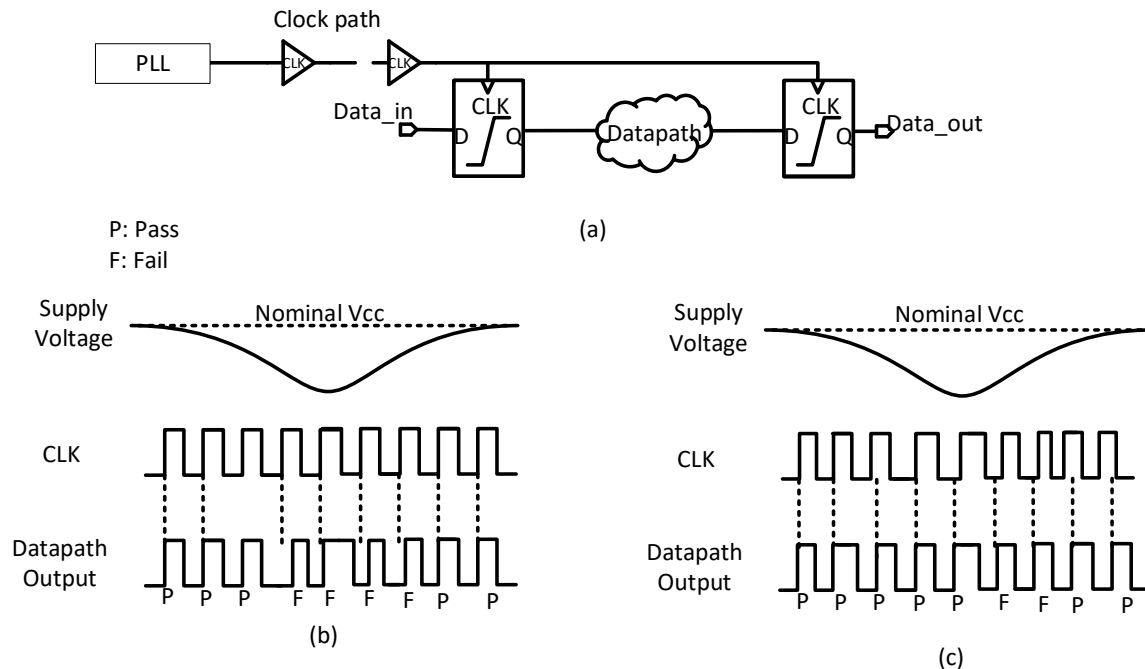
Figure 2.12: (a) Plot of the beneficial jitter effect; (b) constant-period clock; (c) real clock.

Conventional analysis only focuses on the increase in data path delay in the presence of supply noise as shown in Figure 2.12 (b). However, in reality, the clock path also sees a noisy supply, which causes the clock period to gradually stretch during supply downswings (or compression during supply upswings). This clock period modulation effect results in an extra timing margin that compensates for the slowdown in the data path as shown in Figure 2.12(c). Figure 2.13 illustrates how the compensation effect improves the setup time margin. In the presence of supply noise, the maximum data path delay occurs when the supply voltage is at its lowest point, denoted as "A". The corresponding clock edge (i.e., the first edge) which triggers the longest data path delay signal is launched from the clock source at a certain point in time before "A" as it has to traverse through the clock path. The second edge, which will eventually sample the longest delay signal, is launched one clock period after the first edge. It experiences a lower average supply voltage due to the supply downswing, and thus takes a longer time to propagate through the clock path. This makes the clock period longer, compensating for the increased data path delay.
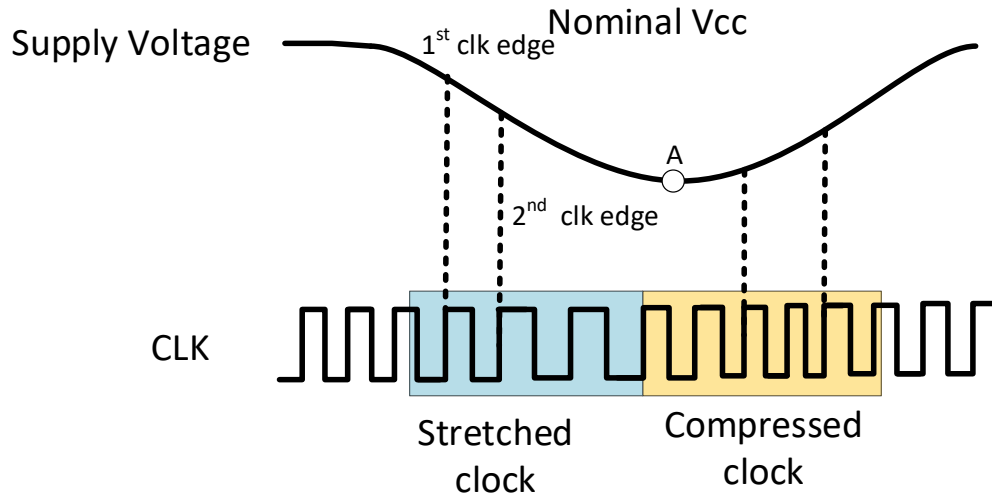
Figure 2.13: Ilustration of clock waveform under resonant supply noise

When it comes to talk about timing analysis, usually we use the word "slack", which is a metric to know if the path is having any failure or not. Usually the "slack: can be expressed by the clock period T_{CLK} minus the actual data path delay T_{DATA}.

$$slack = T_{CLK} - T_{DATA} \qquad (2.9)$$

If the slack is positive, that means there is no failure.

Dong Jiao et al., in [10], present how the slack of a timing path can vary depending of the sensitivity of the clock path. Basically, the sensitivity of the clock and data paths is defined by the normalized change in its delay respecting to the normalized change of the voltage:

$$S_{clk} = \frac{\%increase\ in\ delay}{\%decrease\ in\ supply} \qquad (2.10)$$

They present a detailed HSPICE simulation, which reveals a significant improvement in the worst-case setup time margin even when the clock path and the data path have the same delay sensitivities (previous works present an improvement only when clock sensitivity is higher than data path sensitivity). For a clean clock source, it is clear the slack degradation it is going to be worst as it is shown in Figure 2.14.
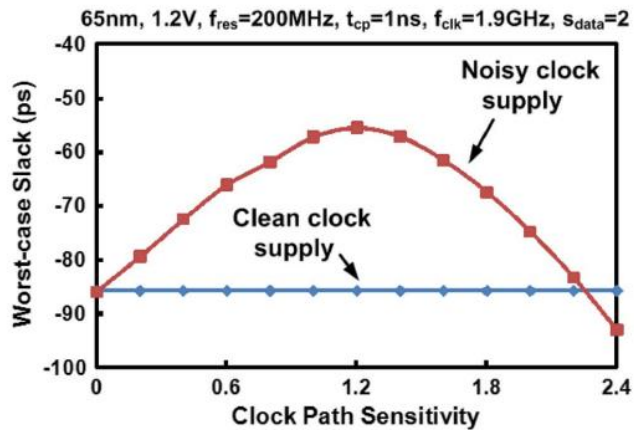
Figure 2.14: Dependency of worst-case slack on clock path delay sensitivity. *From [10]*

Also, their work shows the dependency of worst-case slack on clock path delay, describing an optimum point depending on how long the clock is. Figure 2.15 shows how for extremely long or short clock path delays, the slack considering the beneficial jitter effect (i.e. noisy clock supply) approaches the conventional analysis case (i.e. clean clock supply). This is because a very short clock path makes the clock period modulation effect weaker and conversely, a very long clock path makes each clock edge see a similar average supply voltage.
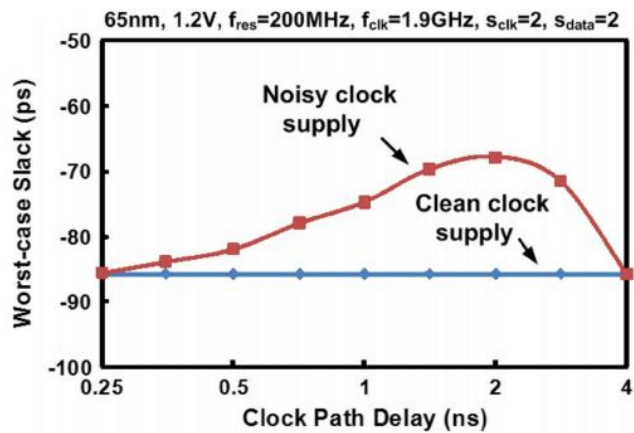
Figure 2.15: Dependency of worst-case slack on clock path delay. *From [10]*

Figure 2.16 shows that at extremely low frequencies, the worst-case slack converges to the clean clock case since two consecutive clock edges see almost the same supply voltage. When the

resonant frequency is high, the noisy clock supply case again converges to the clean supply case. This is because of the negligible difference in the supply voltages seen by two consecutive clock edges due to the averaging effect.
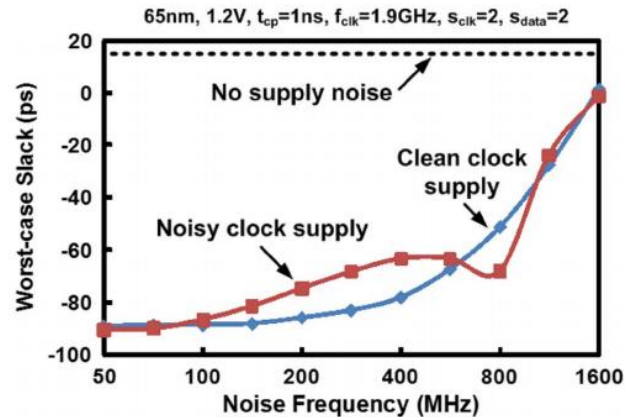
Figure 2.16: Dependency of worst-case slack on supply noise frequency. *From [10]*

Adaptive clocking schemes utilizing this principle of clock-data compensation have been employed in Intel Nehalem processors [16], [20]. In the next section, we are going to explore these different techniques along other that prevent or mitigate errors cause by power supply noise.

2.5.   Noise Mitigation Schemes

Variation has traditionally been handled by margining to ensure a good parametric yield. As variability increases, the growing margins severely degrade the performance and power of a chip. Variation-tolerant designs are becoming more important. [42]

In the design of an integrated circuit, the designer is faced with the challenge of having circuits and systems function over multiple operating points. From the point of view of performance, the circuit must meet its speed requirements over a range of voltages and temperatures that reflect the environment where the circuits are operating. Also, while the performance requirement must be met at a set of worse-case conditions for speed, the power requirement must be simultaneously met at another set of worst-case conditions for power.

Although each design is unique, the resulting instances of fabricated integrated circuits will number potentially in the billions. In addition, the number of components for each of the integrated circuits will also potentially number in the billions. Every single one of the billions of transistors in every one of the billions of circuits is unique. The success of an integrated circuit design is simple measured by the percentage of the fabricated integrated circuits with the transistors, as well as interconnections, meeting all the requirements.

The use of adaptive techniques allows an integrated circuit to adapt for variations in the environment as reflected by both voltage and temperature and also for variations in the fabricated transistors. Adaptive techniques are intended to allow minimization of both dynamic and leakage power, to increase the frequency of operation of the integrated circuit as well. The goal here is to focus in the random variations that affect integrated circuits instead of systematic variations; this is why adaptive techniques such as *Adaptive body bias* (ABB) are not going to be consider later on.

Addressing the issue of excessive margins requires a fundamental departure from the traditional technique of operating every dice at a single, statically determined operating point. Adaptive design techniques seek to mitigate excessive margining by dynamically adjusting system parameters (voltage and frequency) to account for variations in environmental conditions and silicon grade. Thus, a significant portion of worst-case safety margins is eliminated, leading to improved energy efficiency and performance over traditional methods. Broadly speaking, adaptive techniques can be divided into two main categories: "always-correct" techniques and "error detection and correction" techniques. Such approaches rely on scaling system parameters to the point of failure. Computation correctness is ensured by detecting timing errors and suitably recovering from them.

Next, we are going to discuss some of these techniques in order to give more of background regarding the different noise mitigation schemes. The main focus in these next sections is to go over the "Error detection techniques", since most of the techniques used in this work are under this category.

2.5.1. "Always-correct" techniques

The key idea of "always-correct" techniques is to predict the point of failure for a die and to tune system parameters to operate near this predicted point. Typically, safety margins are added to the predicted failure point to guarantee computational correctness. The "always-correct" techniques predict the operational point where the critical path fails to meet timing and to guarantee correctness by adding safety margins to the predicted failure point. The conventional approach toward predicting this point of failure is to use either a look-up table or the so-called *canary* circuits.

Look-up Table-Based approach
In the look-up table based approach, the maximum obtainable frequency of the processor is characterized for a given supply voltage. The voltage frequency pairs are obtained by performing traditional timing analysis on the processor. Typically, the operating frequency is decided based on the deadline under which a given task needs to be completed. Accordingly, the supply voltage corresponding to the frequency requirement is "dialed in". The look-up table approach is able to exploit periods of low CPU utilization by dynamically scaling voltage and frequency, thereby leading to energy savings. Furthermore, owing to its relative simplicity, this approach can be easily deployed in the field. However, its reliance on conventional timing analysis performed at the combination of worst-case process, voltage and temperature corners implies that none of the safety margins due to uncertainties are eliminated.

Canary Circuits-Based Approach

This approach builds a delay chain that mimics the worst-case path on the chip and to use that delay to set the operating frequency. This is called a canary circuit: in the same way that miners sent a canary into the tunnel to see if the air is safe to breathe, the chip uses the canary circuit to determine the frequency that is safe to operate. The canary circuit tracks with the processing and environmental corners, so some of the margin can be eliminated. However, it is still subject to random variations, process tilt, within-die voltage and temperature variations, and other mismatches between the canary circuit and the true critical paths. Characterizing all of these mismatch sources is difficult, so a conservative designer will provide additional margin for the uncertainty. Better yet, the amount of margin can be adjusted at runtime to ensure the part will function at some speed.

There are several systems reported in literature based on canary circuits. One approach uses the replica path as a delay reference for a voltage controlled oscillator (VCO) unit. The VCO monitors the delay through the chain at a given supply voltage and scales the operating frequency to the point of failure of the replica path. An example of such approach is Uht's TEATime in Figure 2.17 [44].

A toggle flip-flop initiates a new transition through the replica path every cycle. The transition is correctly captured at the receiving flip-flop only if the clock period is greater than the propagation delay through the replica path. A simple up-down counter is used to control the VCO frequency output via a digital-to-analog converter (DAC).
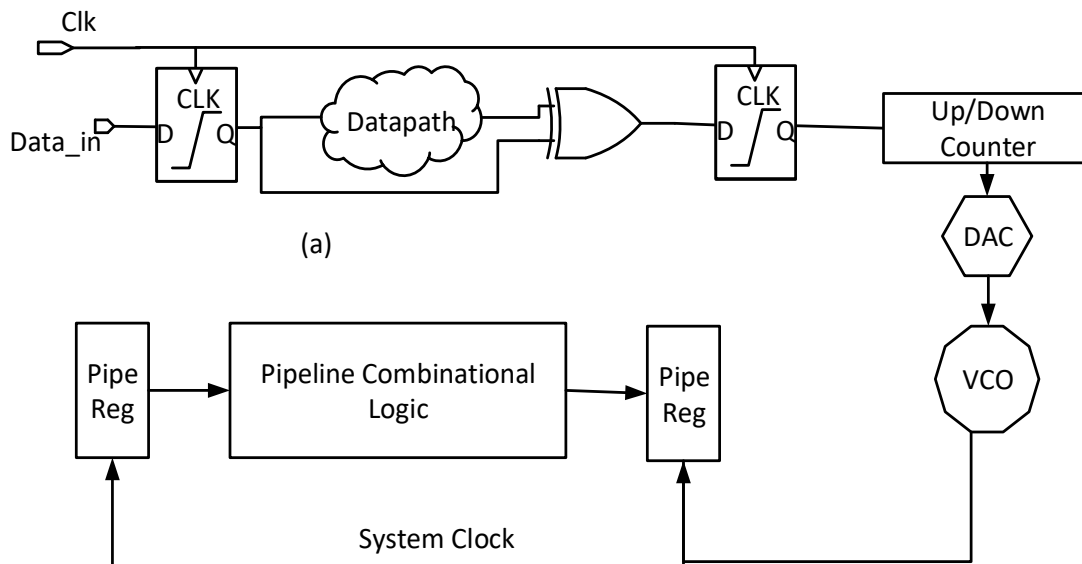
Figure 2.17: Uht's TEATime: A canary circuit-based approach.

Micro-architectural Techniques

A potential shortcoming of both techniques mentioned above is that they seek to track variations in the critical-path delay and, consequently, cannot adapt to input vector-dependent delay variations. The processor voltage and frequency are limited by the worst case path, even if it is not being sensitized. This issue is addressed by several micro-architectural techniques discussed in literature, specifically related to adder architectures [50] [51]. Such designs exploit the fact that

the worst-case carry-chain length is rarely sensitized. This allows them to operate the adder block at the higher frequency than what is dictated by the worst-case carry path. If a latency-intensive add operation is detected, then the clock frequency is halved to allow it to complete without errors. An example of such a design is the stutter adder [50] which uses a low-overhead circuit for a priori determination of the carry-chain length. If the carry-chain length in a cycle exceeds a certain number of bits, then a "stutter" signal is raised which clock-gates the cycle. Thus a "long" adder computation is effectively given two cycles to execute. The authors report that 95% of cases, the adder required only one cycle to compute. Lu [50] proposes a similar technique where an "approximate" but faster implementation of a functional unit is used in conjunction with a slow but always-correct checker to exploit typical latencies and clock the system in a higher rate.

2.5.2. <u>Error Detection techniques</u>

The key concept of these schemes is to scale the system parameters (e.g., voltage and frequency) until the point where the processor fails to meet timing. A detection block flags the occurrence of a timing error after which a correction block is engaged to recover the correct state. To ensure that the system does not face persistent errors, an additional controller monitors the error rate and tunes voltage and frequency to achieve a targeted error rate.

Allowing the processor to fail and then recover helps eliminate worst-case safety margins. This enables significantly greater performance and energy efficiency over "always-correct" techniques. Furthermore, by tuning system parameters based on the error rate, it is possible to exploit the input vector dependence of delay as well. Instead of safety margins, such systems rely on successful detection and correction of timing errors to guarantee computational correctness. The net energy consumption of the system is essentially a trade-off between the increased efficiency afforded by the elimination of margins and the additional overhead of recovery. Of course, the overhead of recovery can make sustaining a high error rate counterproductive. Hence, these systems typically rely on restricting operation to low error rate regimes to maximize energy efficiency.

Variability and Timing Monitors
The increasing timing uncertainty due to noise processes as technology scales is creating a need for on-chip timing sensors that can be explained using Figure 2.18. During the early design stages of the integrated circuit, assuming it is a microprocessor, an iterative process is used to develop the architecture to meet the performance targets of the intended application.
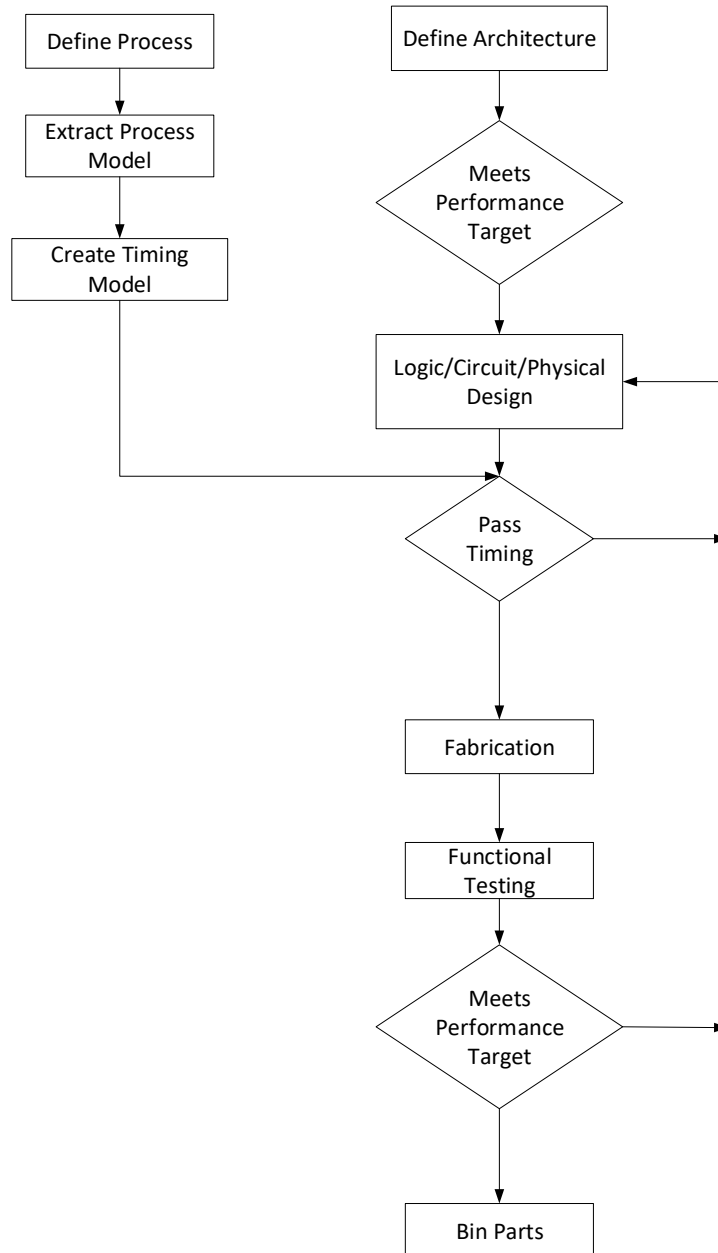
Figure 2.18: A simplified design flow with emphasis on the development of the target performance and testing to ensure that performance is met.

Once the architecture is defined, the microprocessor passes through logic, circuit, and physical design. Models describing the timing of the target technology are used to predict the timing of the microprocessor during its design phases. Once timing is met, the processor is fabricated and tested. If performance targets are met, the microprocessor will be binned into performance categories and sold. If not, the design cycle must iterate at some point to fix the errors. When the timing models can accurately predict the performance of the microprocessor, even when within-die variation is significant, adding a de-sign margin and binning is sufficient for determining the performance of the microprocessor [47]. However, as sensitivity to environmental conditions increases, the needed margins to ensure functionality cause valuable performance to be lost. Because much of the timing

variation (caused by such things as power supply noise and temperature shifts) is related to workload, it can be considered systematic noise and compensated for using dynamic voltage and frequency scaling (DVFS).

DVFS is typically used to optimize the power/performance of a micro-processor, but if the DVFS system can sense a change in temperature, workload, etc.; then, it can compensate for environmental noise and recover some of the design margin. For DVFS to be functional, it must have a means to determine the operating point of the micro-processor. This can be done using workload estimates and look-up tables, but this is usually expensive in terms of calibration time and complexity. Another solution, especially when dealing with fast environmental changes like supply voltage noise, is to use on-chip sensors to monitor the operating condition. Such sensors, typically called critical path monitors, can provide real-time performance information to DVFS systems with a simpler calibration.

The critical path is used because it is the benchmark of timing and is most sensitive to environmental conditions. In addition to providing real-time timing analysis, critical path monitors are extremely useful as an aid in testing microprocessors. Since there is a cost overhead to including critical path monitors, they must provide better performance than just binning and margining by themselves.

In order to build an effective critical path monitor, it is essential to under-stand the sensitivity of path delay to noise. The typical logic path begins at a latch and ends at a latch: on receipt of a clock signal, the data is passed through the logic from the source latch to the final latch. SRAM critical paths are more complicated than logic paths because the control signal often crosses supply voltage boundaries and interfaces with analog sense-amps. Because of this, we will ignore the intricacies of SRAM and just deal with the timing of regular logic.

Critical path monitors are generally used as part of a closed loop DVFS control system. A number of critical path monitors in association with DVFS systems have been reported in the literature [48] [49]. While the specific details of the implementations vary, they all share a basic structure similar to the block diagram shown in Figure 2.8.

Razor sequential circuits
Razor is a circuit-level timing speculation technique based on dynamic detection and correction of speed path failures in digital designs. The main path through the sequential element is unchanged, but a secondary checking path samples the input slightly later. If the two results agree, the circuit is operating correctly. If they differ, the data missed its setup time at the main path but made it for the later sampler, so the frequency is slightly too high or the voltage is slightly too low. This error is reported to a system controller. If the system is designed with a replay mechanism to repeat operations from a last known good state, the operation can be repeated at a lower frequency or higher voltage where it works correctly. Thus, with Razor, it is possible to tune the supply voltage to the level where first delay errors happen. In addition, voltage can also be scaled below this first point of failure into the sub-critical regime, thereby deliberately tolerating a targeted error rate. Due to the strong data dependence of circuit delay, only a few critical instructions are expected to fail while the majority of the instructions will operate correctly.

Figure 2.19 shows the basic concept of the Razor flip-flop [52] [53]. The main path uses an ordinary flip-flop, while the checking path uses a latch. The flip-flop samples on the rising edge

of φ_p while the latch samples some time later on the falling edge of Kp. Figure 2.19(b) illustrates the operation of the circuit. If the data arrives at least a setup time before the rising edge of φ_p, both elements sample the same value. If the data arrives late, the flip-flop misses the data and the XOR generates an ERR signal. The ERR signals from all the flip-flops in the system (or at least those on potentially critical paths) are ORed together to indicate an error and trigger the replay mechanism.
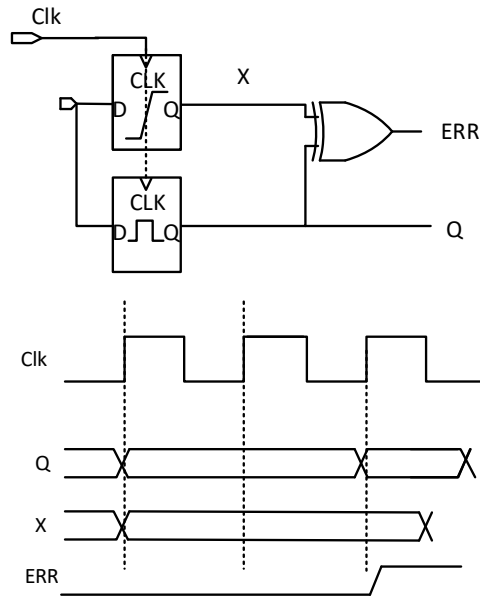
Figure 2.19: Example of a Razor sequential circuits

The operating voltage and frequency are adjusted until the system is barely working so that very little margin is provided: the circuit is functioning "on the razor's edge." Variations such as power supply noise, unusually large crosstalk, or even activation of rarely triggered critical path, are sufficient to delay the arrival of D and cause an occasional error. The width of the clock pulse presents a trade-off between error detection and hold time. Wider pulses allow later inputs to be detected as errors, which increases the allowable difference between typical and worst-case delay. However, the hold time increases with the pulse width, just like a pulsed latch. Managing long hold times is difficult, so a relatively narrow pulse (e.g., < 3 FO4 delays) is preferable.

The Razor circuit has the drawback that the flip-flop may become metastable if D changes during the aperture. If Q resolves to the same value as the latch, no error will be flagged, but the propagation through the flip-flop can increase by an unbounded amount of time. Reference [52] suggests a adding a meta-stability detector, which significantly increases the overhead of the circuit.

Adaptive Clock Distribution
Several adaptive circuit techniques have been reported in the literature aiming to reduce the effect of the voltage drops by explicitly sensing the variation with on-die monitors and adjusting the operating condition (e.g., clock frequency). Although this is effective at low frequencies, the

chances to mitigate high-frequency drop are very limited. Also, techniques like resilient timing-error detection and recovery circuits can be very useful since they detect the timing violation, isolates the error from corrupting the architecture state, and correct the error through instruction replay [2]–[5]. A resilient design is highly effective at mitigating the impact of high-frequency drops on performance. However, the architectural design complexity for implementing error recovery into a high-performance microprocessor while ensuring coverage for all failure scenarios is a significant challenge.

Finally, adaptive clock distribution (ACD) along with clock gating techniques has become an interesting solution to mitigate the effect of voltage drop on microprocessor performance. These techniques take advantage of the clock-data compensation effect, in which both clock and data are affected by the drop, so changing the clock signal compensate changes on data paths. Nevertheless, clock gating can bring synchronization problems between blocks in a high performance microprocessor.

Adaptive clocking schemes utilizing the principle of clock-data compensation have been employed in Intel Nehalem processors [16], [20]. There, a PLL based clock generator is designed to track the supply noise so that the clock period stretching effect is maximized. An alternative way to enhance the beneficial jitter effect is shifting the phase of the supply noise seen by the clock path [17], [18], [19], for example by using an *RC* filtered supply voltage for the entire clock path. A similar approach has been used in Intel Pentium4 processors where the supply noise of the clock buffer is reduced by using a local *RC* filter [21]. However, these existing designs require a large area overhead due to the small resistance and large capacitance requirements.

In [10], authors propose a novel phase-shifted clock buffer design that can improve the maximum operating frequency by 8–27% while saving 85% of the clock buffer area compared to previous phase-shifted designs for a typical resonant frequency range between 100 MHz and 300 MHz. They proposed design can be used in conjunction with adaptive clocking schemes for further improvement in chip performance.

Other example of Adaptive clock Distribution (ADC) was published by Bowman et al. [1]. The design integrates a tunable-length delay prior to the global clock distribution to prolong the clock-data delay compensation in critical paths during a drop. The tunable-length delay (TDL) prevents critical path timing-margin degradation for multiple cycles after the drop occurs, thus allowing a sufficient response time for dynamic adaptation. An on-die dynamic variation monitor detects the onset of the drop to proactively gate the clock at the end of the tunable-length delay to eliminate the clock edges that would otherwise degrade critical-path timing margin.

As illustrated in Figure 2.20, the adaptive clock distribution design contains a tunable-length delay, the DVM_{ROOT}, and a clock-gating circuit. The salient feature is the tunable-length delay between the clock generator and the global clock distribution. The tunable-length delay, consisting of scan-programmable transistor and interconnect delay components, prolongs the path clock-data delay compensation during a voltage drop by extending the delay and changing the delay sensitivity to V_{CC} in the clock distribution. The tunable-length delay prevents path timing-margin degradation for multiple cycles after the voltage drop occurs to enable a sufficient response time for dynamic adaptation.

Post-silicon calibration aims to achieve a target tunable-length delay and a target delay sensitivity to for the clock distribution. Since the adaptive clock gating occurs after the tunable-length-delay circuit, the target tunable-length delay equals the product of and the number of cycles for an adaptive response. Performing several fmax tests, data path sensitivity can be found. This calculations sets target for clock distribution.

In comparison to a conventional clock distribution, silicon measurements from a 22 nm test chip demonstrate simultaneous throughput gains and energy reductions of 14% and 3% at 1.0 V, 18% and 5% at 0.8 V, and 31% and 15% at 0.6 V, respectively, for a 10% V_{CC} drop
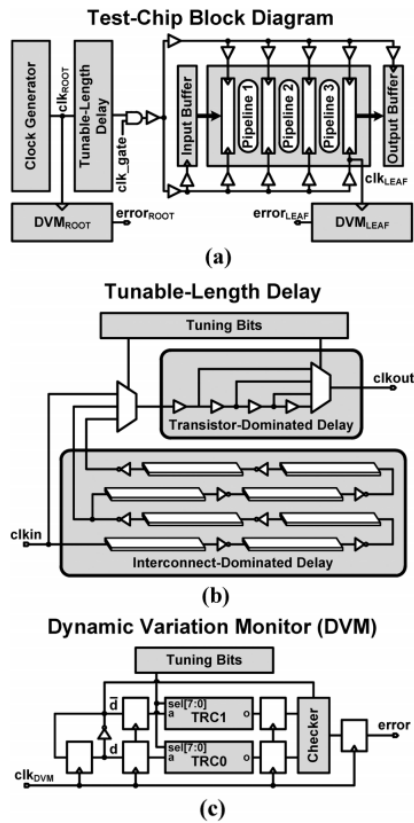
Figure 2.20: (a) Test-chip block diagram of the all-digital dynamically adaptive clock distribution integrated into a 3-stage pipeline circuit. The adaptive clock distribution consists of (b) a tunable-length delay, (c) a dynamic variation monitor at the clock root . From [1]

Dynamic Voltage-Frequency Scaling

Besides technology scaling, one of the most effective ways to reduce active power consumption is by lowering Vcc. Ideally, quadratic power savings are observed as displayed in Figure 2.21, since dynamic power is represented as follows:

$$P_{dyn} = f * C_L * V_{cc}^2 \tag{2.11}$$

where C_L is the capacitive load, f the operating frequency and V_{cc} the operating voltage.

Vcc reduction can be applied to a complete chip, but it is most effective when it is applied to local voltage domains with own performance requirements. A common approach is to perform dynamic supply scaling, which exploits the temporal domain to optimize Vcc at runtime. This technique dynamically varies both operating frequency and supply voltage in response to workload demands. In this way, a processing unit always operates at the desired performance level while consuming the minimal amount of power. Two basic flavors exist, namely adaptive voltage scaling (AVS) and dynamic voltage scaling (DVS). AVS is a closed-loop approach, and its operating points are based only on the frequency. Software decides on the performance required for the existing workload and selects a target frequency. The voltage is then automatically adjusted to support this frequency. AVS is considered as the most effective technique for achieving power savings through Vcc scaling.
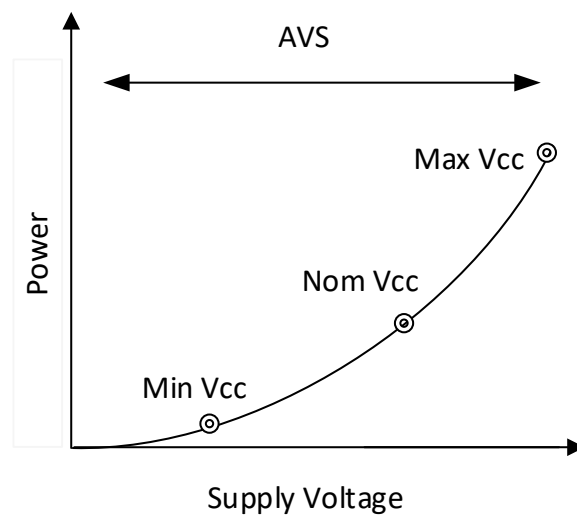
Figure 2.21: Power trends as a function of the supply voltage

Having a dedicated voltage domain per each functional block can improve energy efficiency by enabling AVS. In general, advanced CMOS devices cannot directly handle the voltage levels provided by either a power supply unit or a lithium-ion battery, mandating off-chip voltage regulators. While off-chip switching regulators can offer one step conversion from the sources with 90% efficiencies, they increase the number of the off-chip components and the package bumps. Particularly, since package bump pitches scale at much slower rate than device scaling, the implantation overhead associated with the bumps continues to scale up [22].

Therefore, the total number of the voltage domains driven by off-chip regulators is often restricted. This creates a "shared rail" scenario where multiple functional blocks share a common supply voltage. Given these drawbacks of off-chip DC-DC converters, there has been a surge of interest in building on-chip converters to implement multiple on-chip voltage domains. Each functional block has its own set of power gates allowing it to form an independent power island that can be turned on or off as needed to cut down leakage power during idle periods as shown in Figure 2.22(a). On the other hand, as shown in Figure 2.22 (b), if the power gates can be transformed into integrated voltage regulators (IVR) without imposing substantial implementation overhead

such as extra passive components and additional area overhead required for the control block, a cost-effective AVS can be achieved through fine-grain voltage domains. [22]
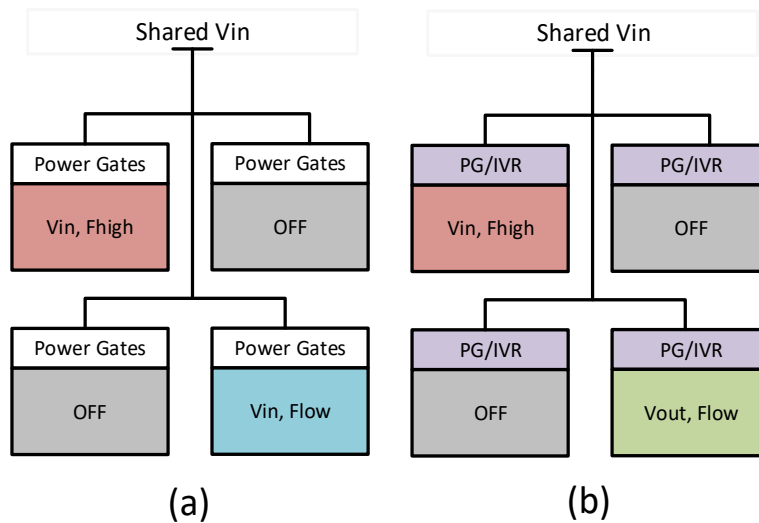
Figure 2.22: Multiple cores forming (a) independent power islands with power gates and (b) independent fine grain voltage domain with IVRs.

Alternatively, DVS is an open-loop approach, and it is based on the selection of operating points from a predefined {f,V} table. DVS is one of the popular approaches in power reduction. Vcc is dynamically lowered to an extent where required performance of the target system is ensured. Significant power reduction is possible with DVS, since dynamic power of CMOS circuits is proportional to the square of Vcc. This method allows to achieve energy efficiency in systems that have widely variant performance demands. As Vcc decreases, transistor drive currents decrease, bringing down the speed of operation of a circuit. A DVS system adjusts the supply voltage, operating the circuit at just enough voltage to meet performance, thereby achieving overall savings in total power consumed.

A detailed study of processor workloads shows that high levels of activity fluctuation exist at finer timescales that traditional AVS cannot track [23]. Also, to maximize power savings from fine grain voltage domains supporting high performance modern microprocessor workloads, the corresponding IVR should be able to support a wide dynamic voltage scaling (DVS) range, from the near-threshold voltage (NTV) region to the turbo region with high efficiency across the whole range.

While static techniques such as clock tuning, adaptive body bias, and adaptive supply voltage can effectively compensate process variations, other variations such as temperature, voltage drops, noise, and transistor aging are dynamic and change throughout the lifetime of the processor. These cannot be compensated using a static technique and are typically guardbanded using either reduced frequency or higher supply voltage. This guardbanding is expensive in terms of performance and power and is becoming prohibitive as design margins shrink. To achieve an energy-efficient microprocessor which operates correctly in the presence of these variations, a method of sensing the environment and responding by changing voltage, body bias, or frequency is necessary. In this section, we describe one implementation of a dynamic adaptive processor design [39]. In this section, several implementations of a dynamic adaptive voltage scaling are presented.

A test-chip in 90nm CMOS technology (Figure 2.23) contains a TCP offload accelerator core, a data input buffer, VCC drop sensors, thermal sensors, a dynamic adaptive biasing (DAB) control unit, distributed noise injectors, body bias generators, and a three-PLL dynamic clocking unit [44]. The DAB controller receives inputs from the thermal sensors and drop detectors. Average supply current is sensed by the off-chip voltage regulator module (VRM) and digitally communicated to the DAB controller on chip. The programmable noise injectors are used to generate various supply noises and load currents, in addition to that generated by the core during normal operation. The DAB controller drives the dynamic frequency unit, body bias generators, and voltage setting of the off-chip VRM to dynamically adapt frequency, body bias, and VCC to achieve optimum settings for the given conditions.
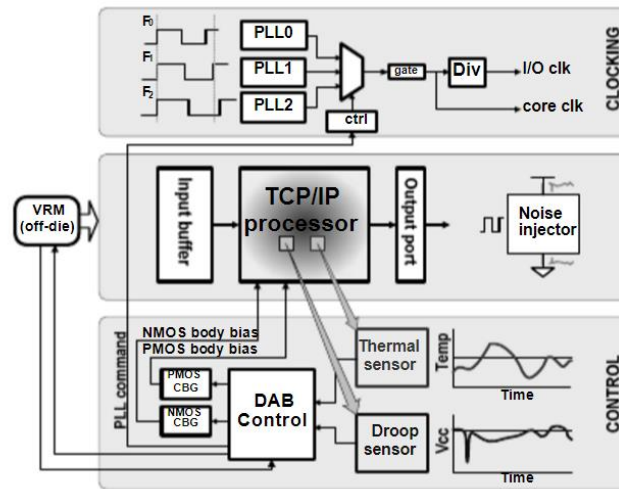
Figure 2.23: Block diagram of the dynamic adaptive TCP/IP processor. *From [44]*

The control is designed to be fast enough to respond to second and third drops in voltage as well as changes in temperature and overall chip activity factor. Responding to the relatively fast VCC drops also requires a method for changing frequency quickly without waiting for a PLL to relock. The clocking subsystem, contains three PLLs running at independent frequencies and a multiplexer to select between them in a single cycle while ensuring that there are no shortened clock cycles. The dynamic frequency algorithms are implemented in the DAB control, and commands are sent to the PLL block to switch between PLLs and update PLL divider values.

This combination reduces the guardband needed for maximum temperature and, in this example, results in a 1.4% increase in average frequency over the duration of the test. In a similar way, dynamic response to voltage drops allows the drop guardband to be reduced or removed, resulting in increased average performance or reduced power consumption—in this instance, increasing the average frequency by 5%.

In previous publications like [22], a digitally-controlled fully integrated voltage regulator (IVR) enables wide autonomous DVFS in a 22 nm graphics execution core (Figure 2.24).
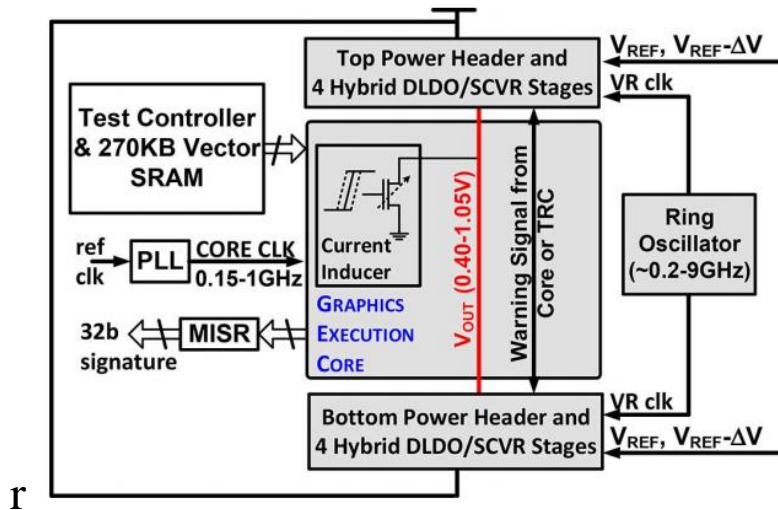
Figure 2.24: Block diagram of a digitally-controlled fully integrated voltage regulator (IVR) enables wide autonomous DVFS in a 22 nm graphics execution core

Part of the original power header is converted into a hybrid power stage to support digital low-dropout (DLDO), and switched-capacitor voltage regulator (SCVR) modes, in addition to the original bypass and sleep modes. Inductor-based switching converters offer inherently higher efficiency over a wide output voltage range than linear regulators. However, implementing high quality inductors on CMOS technology is challenging, the overall silicon footprint required for the proper package inductor hookup is still large enough forcing multiple functional blocks to co-exist with a shared supply rail. LDO has been presents good area efficiency and ability to support high power density loads. However, the performance of a conventional LDO is heavily bounded by the error amplifier used in the feedback.

Chapter III

Clock-Data Compensation Effect Studies

This chapter corresponds to the construction of a baseline design and to evaluate the effect of the clock-data compensation (CDC) when a voltage drop arises. The main focus is on high-frequency voltage drops, since available literature on this topic show that the resonant supply noise typically lies in the 50-300MHz range with a maximum magnitude of approximately 10% of Vcc for voltage drops [19]. This chapter evaluates the impact of the clock tree length on the critical data path when the supply voltage is changing. For this, several test-cases are proposed using different clock lengths. Also, for a deeper exploration of the CDC effect, these baseline studies take into account the behavior of the critical path under different process corners.

3.1   Powergrid Model Definition

As a first order approximation, the noise component in the power supply is caused by the resonance between the package inductance and die capacitance [17], making the supply voltage behavior similar to a RLC circuit.

A change in the load cause this voltage drop and it can be due different factors: a high-power demanding instruction run by the processor, a memory access, a turn on logic that was power/clock gated, etc. At the end, this is translated in a change in the current profile of the load, which in this case contains buffers in series for the data path, clock buffer in the clock tree, along with a receiving and a launching flip-flops.

Modeling the powergrid as a second order circuit will allow capturing PDN variations closer to reality[1]. The configuration used corresponds to a reduced model of Figure 2.3, were board/package/die parasitics are the only one considered, since the magnitude and duration of a voltage drop depend on the interaction of capacitive and   inductive   parasitics   at   the   board, package, and die levels with changes in current demand [1].

Not modeling the powergrid will incur in non-realistic results, since the oscillations in the voltage depends on the intrinsic parasitics associated to the real implementation of the circuit (die + package). The variable load is modeled as a current source with a value that can be defined arbitrarily. The system is described in Figure 3.1, where the load and the variable load are connected to the same power supply, causing that any change in the current consumption will affect the voltage level of the circuits.

---

[1] A detailed model of a PDN is much more complex than that in a real design. The second order model pursue to capture the essence of PDN variations and generate realistic waveforms.
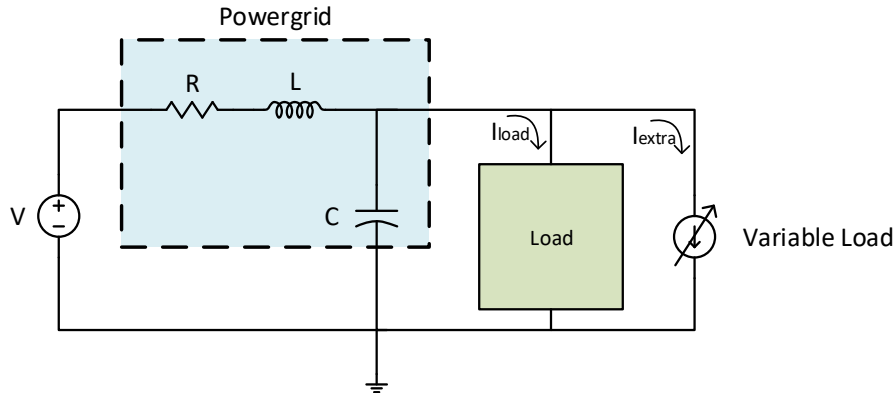
Figure 3.1: Power distribution network model with a variable current source representing the load variability that can generate a voltage drop

The voltage drop is defined with the following characteristics:
- ~10% of Vcc voltage drop.
- Frequency of 100 Hz

In chapter 2, it was explained how the voltage drop is going to depend on many factors. First, it is necessary to understand the current profile of the circuit under study, and define what will be the maximum current consumption.

Figure 3.2 shows the steady-state value of the supply current for the critical path, in which there are some variations due to the switching capacitance of the clock cells. In average for the used load, the current oscillates around the 300 mA value.
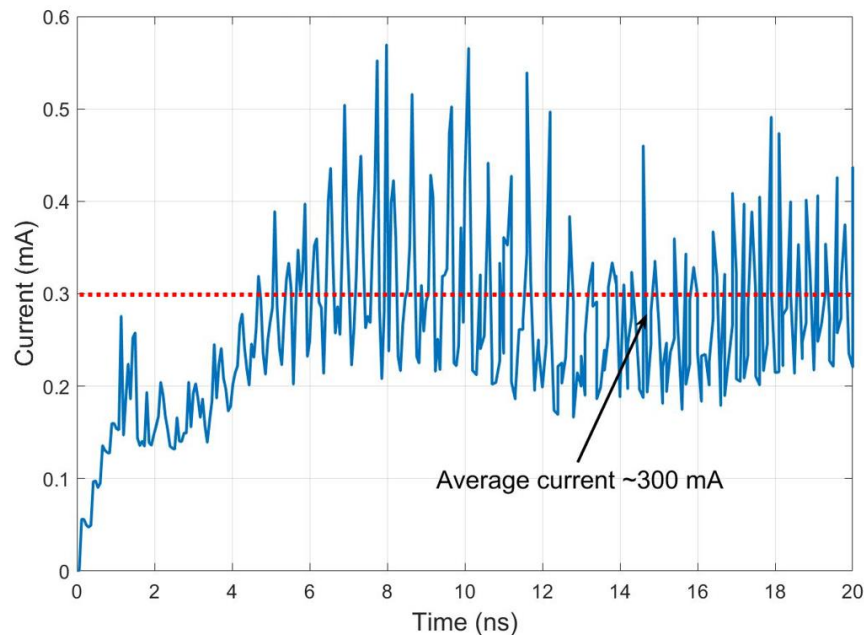
Figure 3.2: Supply current for the critical path waveform oscillating around 300 ua in steady state with variations due to switching activity from clock buffers

The extra load needs to be significant based on the I_{load}, if I_{extra} it is too small will not cause any impact on the voltage waveform. If it is too big can lead to long periods to make the system returns to its steady state. Therefore, I_{extra} is defined to be 0.105 mA in this case, which means an increase of 35% of the load. Note that having a 10% voltage drop does not mean to have 10% increase in the current. [2]

Using the triangle approximation in Eq. 2.3 (Figure 2.2) the value of the voltage drop depends on the following: current peak (I_p), time for the current to reach I_p (T), voltage drop value (v_{nmax}) and the value of inductor (L).

Using these parameters will allow obtaining an estimate of the total inductance in the powergrid L. For T, we are assuming the triangle in Figure 2.2 is symmetric[3].Since the lowest value of the voltage ($V_{max} - v_{nmax}$) will happen 5 ns after the voltage starts to drop, the value of T needs to be halve (2.5 ns). Therefore, we have the following value:

$$v_{nmax} = 0.065V \ (10\% \ voltage \ drop)$$
$$I_{droop} = 0.105 \ mA \ (35\% \ of \ I_{load})$$
$$T = 2.5 \ ns$$

Replacing these values in Eq. 3:

$$v_{nmax} = L * \frac{I_p}{T}$$
$$L = 0.065 * \frac{2.5 \text{x} 10^{-9}}{0.105 \text{x} 10^{-3}} = 1.5 \mu H$$

Based on this value on Eq. 2.8, which gives the value of the capacitor based on the fact we want a resonant frequency of 100 MHz:

$$f_{resonant} = \frac{1}{2\pi\sqrt{LC}}$$
$$C = \frac{1}{L * (2\pi * f_{resonant})^2} = \frac{1}{1.5 \text{x} 10^{-6} * (2\pi * 100 \text{x} 10^6)^2} = 1.69pF$$

Finally, for the ESR value 5 Ω is used.

---

[2] Powergrid non-ideal behavior and many other circuit's parasitics, prevent that to happen and also the voltage drop value is part of the dynamic response of the circuit and not part of its steady state.

[3] Rise time is assumed to be equAl to fall time

3.2 Powergrid Model Response under Variable Load

This section will go through the powergrid model response under a variable load in the system, showing how the voltage supply change when a current step is applied to the second order powergrid model After setting all the values in the powergrid model of Figure 3.1, the obtained response after applying an extra current step [4] is shown in Figure 3.3
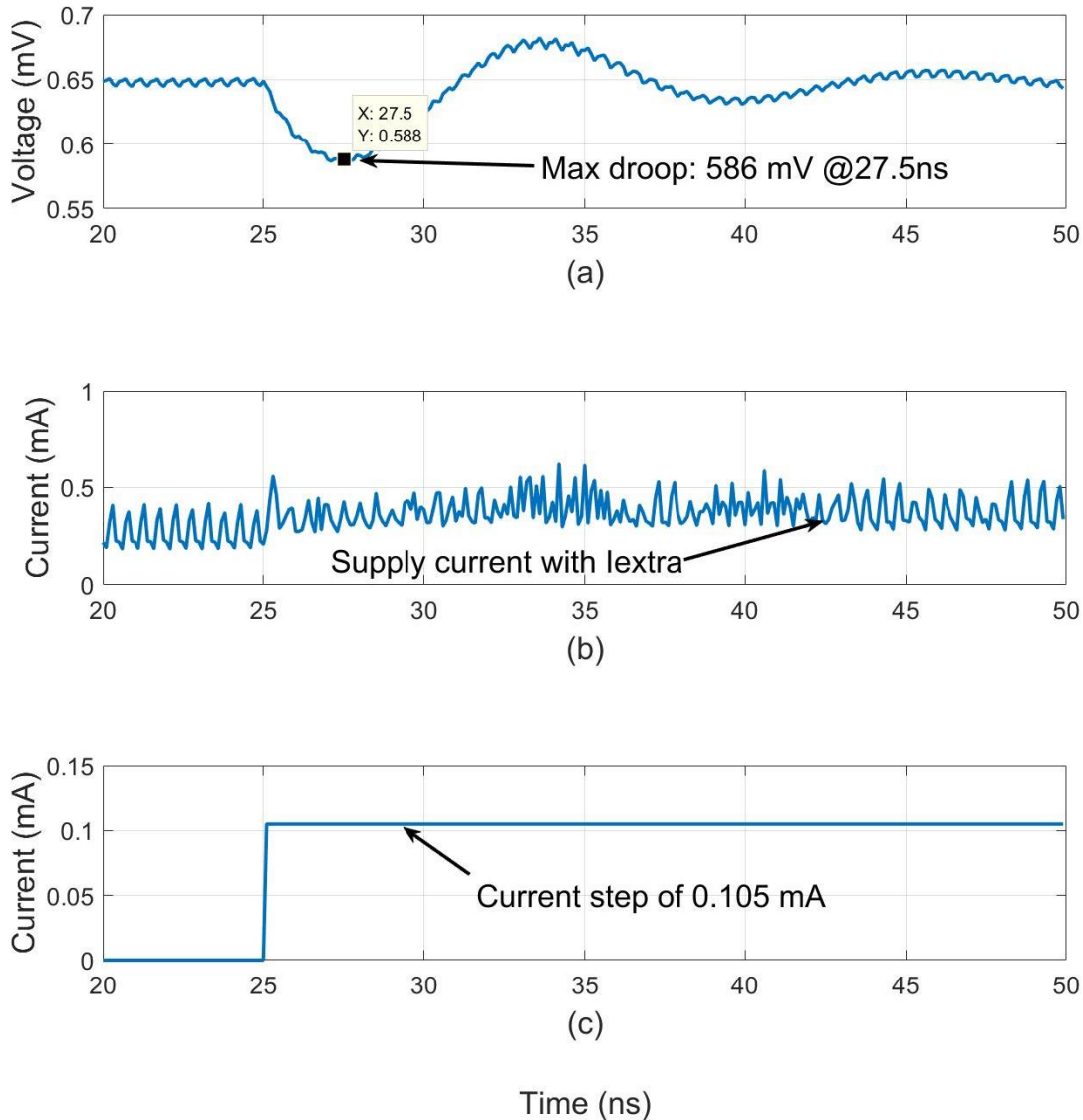
Figure 3.3: Supply voltage response under current step: (a) voltage supply level for data path, (b) supply current when current steps occurs, and (c) current step of 0.1 mA

---

[4]The extra current step represents the load variability that can generate a voltage drop

Figure 3.3 presents the supply voltage response at the time to have a current step in the system. The second order effect can be appreciated in Figure 3.3(a), in which the maximum value of the voltage drop occurs at 27.5ns due to its frequency (100 MHz). The value of the voltage at this pointer is 0.588 mV, which corresponds ~10% voltage drop in the critical path.

For the Figure 3.3(b), the initial current in the circuit that is about 0.3 mA. Once the extra current step happens, the final value of the current is 0.405 mA, in between these two values there is a transient which corresponds to the circuit respond to the impulse at 25 ns. Finally, in Figure 3.3(c) the applied current step is shown.

3.3   Circuit Description

The CDC effect is demonstrated on a circuit which is going to mimic the critical path in a chip. The analysis is done on different scenarios, showing how timing degradation of the data path is being compensated or degraded by the clock period change under the presence of a voltage drop.

First, the operating conditions of the circuit need to be determine:

- Clock Frequency: 2GHz
- Voltage supply (Vcc_exp): 0.65 V

In Figure 3.4, there are two types of circuit topology that are proposed: one circuit with a clock tree and another one without it. The general scheme of the circuits is depicted there.
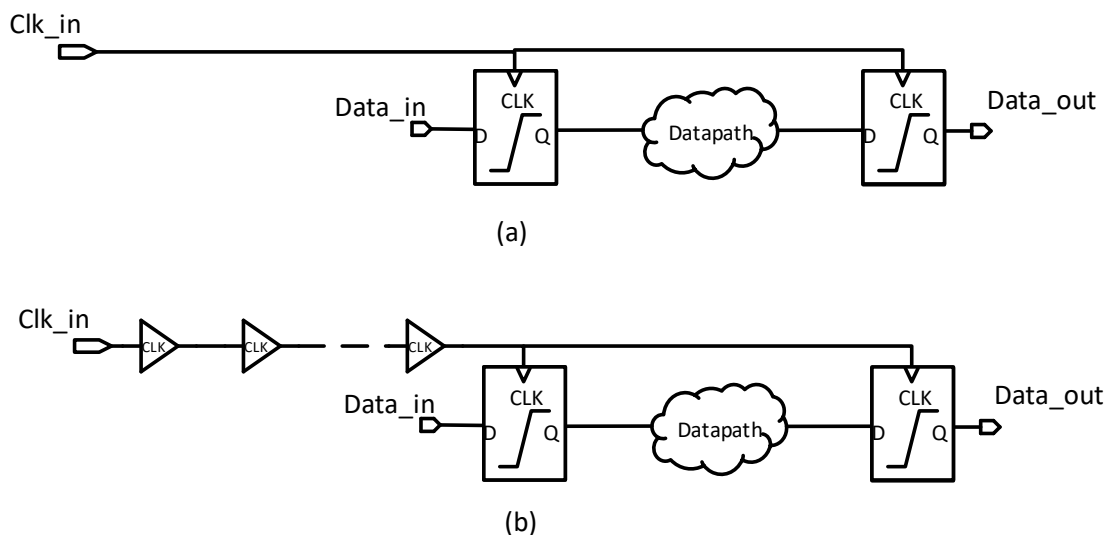
Figure 3.4: Device under test topologies: (a) no clock tree topology; (b)clock tree topology.

The data path and clock tree characteristics are listed in Table 1.

TABLE 3.1: CHARACTERISTICS OF DATA PATH AND CLOCK TREE

Block	Description	Device Type
Data path	Buffers connected in daisy chain	High-Vt devices
Clock tree	Can vary from 0 to 190 clock buffers	Nominal devices

Nominal devices are preferred in the clock cells. Despite of the fact these cells have more activity factor, the clock tree is a very critical part in a design. Hence, designer needs to make sure process variations and random noise events are going to have less impact in a particular portion of a design. In the other hand, high-Vt and ultra-high-Vt devices are preferred in data paths. Data paths are the bigger portion in a block, therefore is the part that consumes more leakage. Also, variations can have a bigger impact on these cells.

In order to verify the circuit for both transitions rise and fall, a toggle flip-flop is being used at the input of the launching flip-flop (Figure 3.5). This will make sure the data is changing on every cycle.
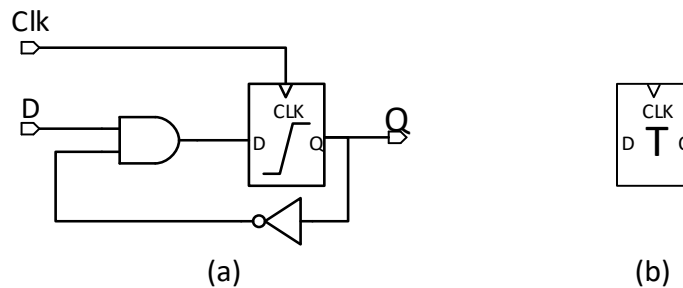
Figure 3.5: Launching configuration: (a) toggle-flip-flop; (b) toggle flip-flop symbol

3.4   Clock-Data Compensation Effect on Critical Path

Based on Figure 3.4, there are three different scenarios that are going to be evaluated as follows:
1. Voltage drop on DUT without clock tree (Test case 1)
2. Voltage drop on DUT with clock tree affecting data path only (Test case 2)
3. Voltage drop on DUT with clock tree affecting both clock and data path (Test case 3)

Test Case 1: Voltage Drop on DUT without Clock tree

This test case will show the impact of the voltage drop on the critical path without a clock tree. An error detection scheme has been implemented in order to detect when the output of the flip-flop

does not have the same value as the input at the time a positive edge of the clock is arriving to the capturing flip-flip. The high-level schematic is depicted in Figure 3.6
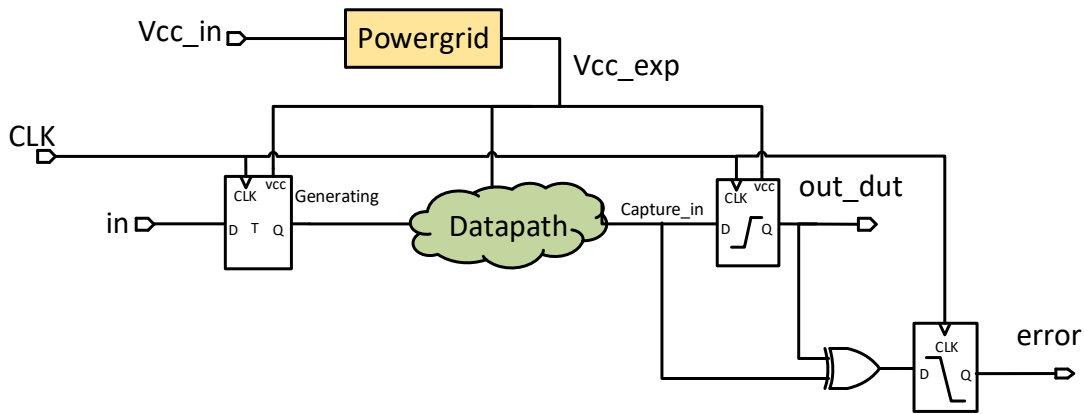
Figure 3.6: Test case 1 voltage drop on DUT without clock tree

In Figure 3.7, for the implemented studied case, it is observed how an extra current demand of around ~120μA (Figure 3.7(b)) is affecting the supply voltage for the data path in the system (Figure 3.7(a)). The voltage drop presents several oscillations, but the most critical one is the first drop called "first-order drop".
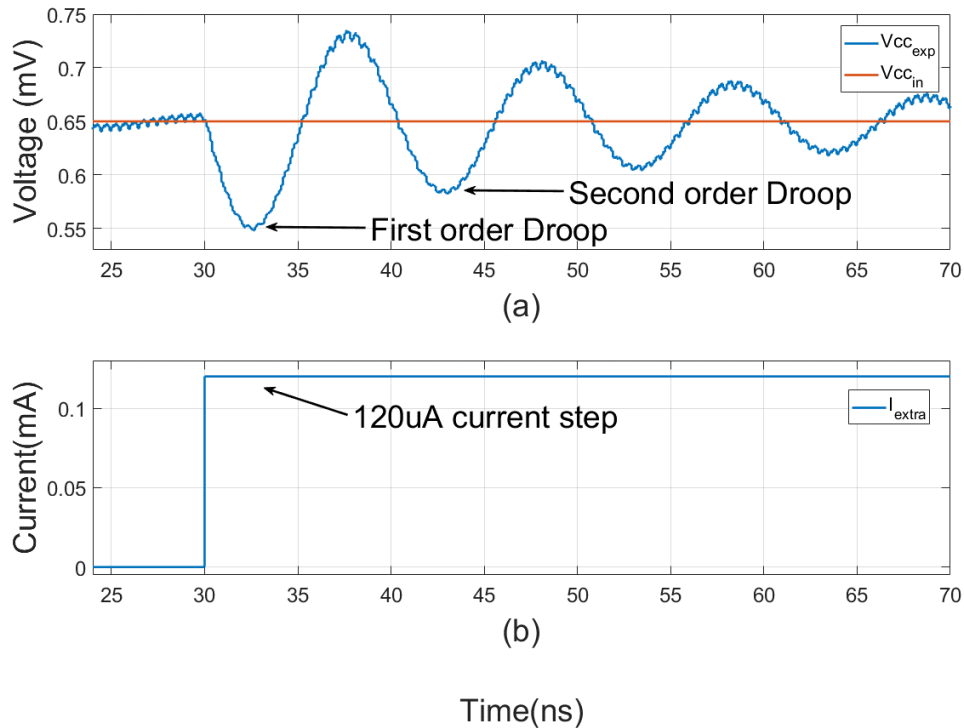
Figure 3.7: Test case 1, supply voltage response under current step: (a) voltage response; (b) current step

The data path has enough slack to afford the voltage drop event. In Figure 3.8, it is shown how the output of the capturing flip-flop follows the input every positive edge of the clock. This is because the system has enough slack to support the current step without having any failure. Also, as it is

expected clock doesn't change neither its frequency nor voltage level. This because there is no clock tree in the system or any device related to the clock for this test case.

When a 140µA current step is applied, the circuit will not have enough slack to support the voltage drop. In Figure 3.9(a), it is observed how the voltage drop causes that the output of the capturing flop is not following its input on every clock cycle, having some missed transitions. In Figure 3.9(b), the error signal is activated due to the missed transitions.
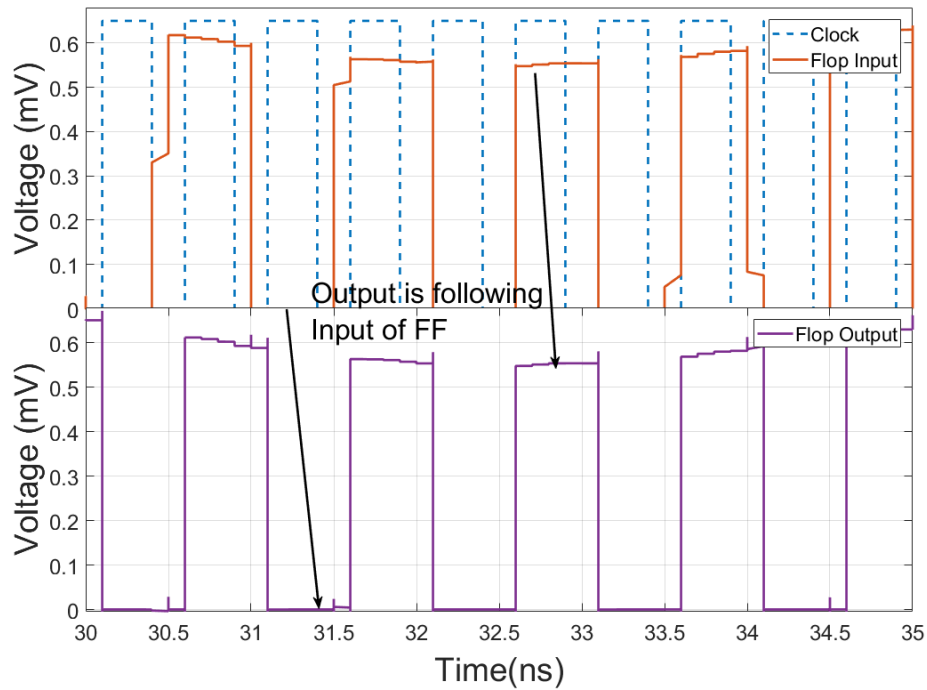
Figure 3.8: Test case 1: Test case 1: input and output of the capturing flip-flop
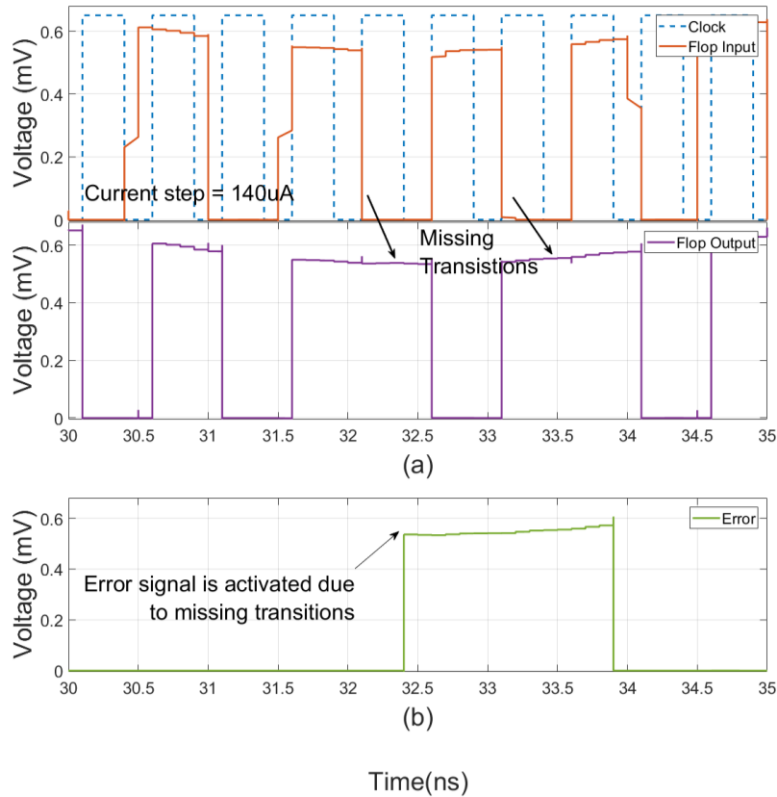
Figure 3.9: Test case 1 circuit response to current step over threshold: (a) In-out of capturing flip-flop; (b) error signal is activated when output of the flip-flop does not match its input

Test Case 2: Voltage Drop on DUT with Clock Tree Affecting Data Path Only

This test case shows the impact of the voltage drop on the critical path with clock tree by applying the same current step as in test case 1 (140µA). This drop will affect only the data path and not the clock tree. The clock tree will be connected to a clean supply. Doing this, it will be shown how the clock will not compensate since is not being affected by a voltage drop. The high-level schematic looks like Figure 3.10.
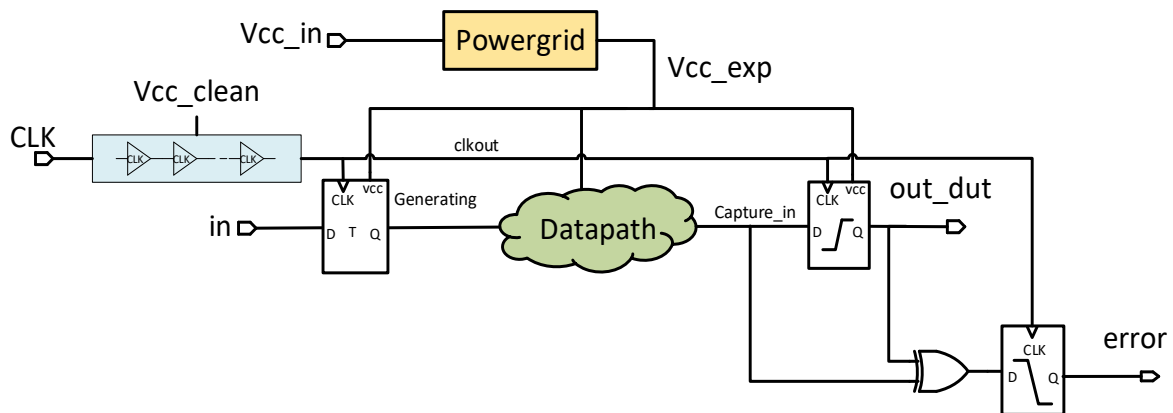
Figure 3.10: Test case 2: voltage drop on DUT with clock tree affecting data path only

The clock frequency and latency. The "clkout" signal corresponds to the clock signal after the clock tree. The latency between these two signals is around 430ps.

In Figure 3.11(a), it is observed how the frequency of the clock is not affected by the voltage drop and still is 500ps. Having said that, it is appreciated how there is no compensation by the clock, and the systems still fails around 32.25 ps, and is missing two transitions. The transitions happen not to be exact the same due to the small clock variations between a clean clock source with sharp edges (test case 1) and a clock source driven by standard cells (test case 2). In other words, Test-case1 and Test-case2 behave the same way under a voltage drop event.
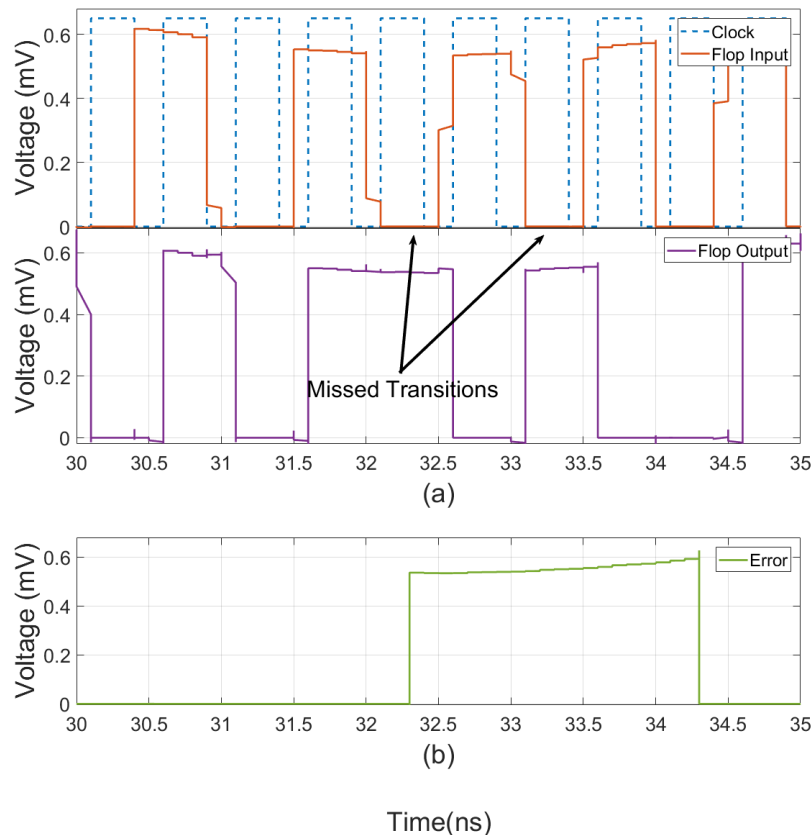
Figure 3.11: Test case 2 circuit response to current step over threshold: (a) In-out of capturing flip-flop; (b) Error signal

Test Case 3: Voltage Drop on DUT with Clock Tree Affecting both Clock and Data Path

This test-case will show the impact of the voltage drop on the critical path with a clock tree, in which the drop is affecting both clock tree and data path. With this scheme, the CDC effect can be observed. Clock tree and data path sensitivity need to be matched, in which the clock can be selected in a way to overcome the voltage drop (at least when the voltage is decreasing). This because on the first part of the drop the clock will stretch, and in the second half (when the voltage is recovering) the clock will compress, which can or not cause a timing failure depending on the data path sensitivity to voltage supply changes.

The high-level schematic looks like the following figure:
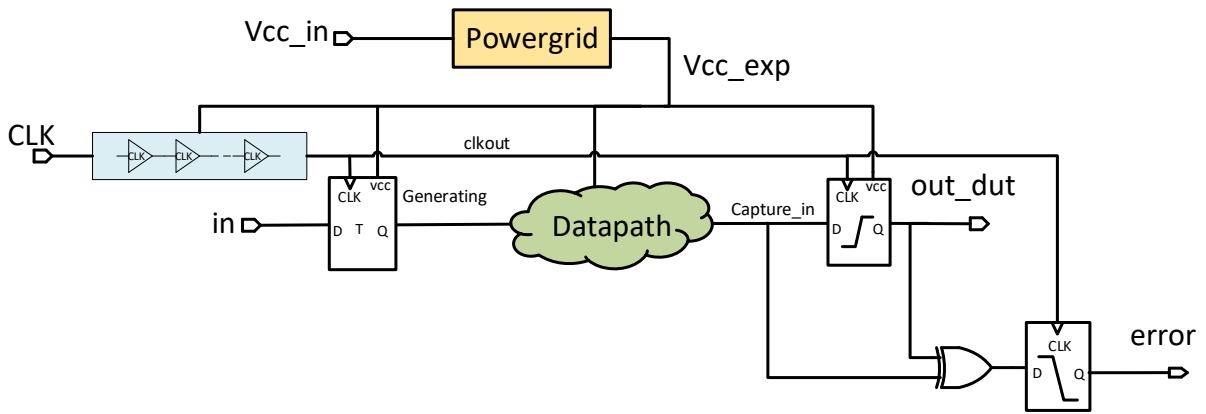
Figure 3.12: Test-case 3: Voltage drop on DUT without clock tree

Figure 3.13(a) shows how after applying a 140µA current step, the output of the flip-flop follows the input of the same on every clock cycle. The same current step caused failures on test-case1 and test-case2 because the clock will stretch when the voltage is decreasing, compensating the delay increase in the data path.
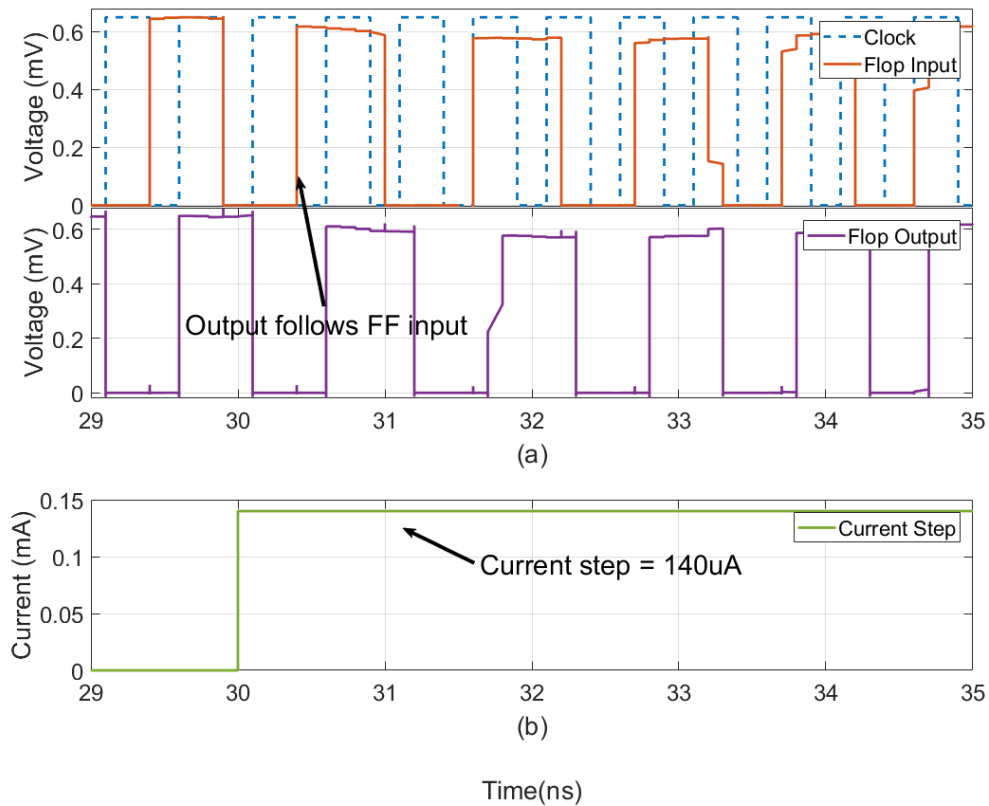
Figure 3.13: Test Case1 Supply voltage response under 140µA extra current step: (a) In-out of capturing flip-flop; (b) Extra current step

3.5 Clock Tree Length Impact on Critical Path

For extremely long or short clock path delays, the slack considering the beneficial jitter effect (i.e. noisy clock supply) approaches the conventional analysis case (i.e. clean clock supply). This is because a very short clock path makes the clock period modulation effect weaker and conversely, a very long clock path makes each clock edge see a similar average supply voltage.

Choosing the clock tree length to have the optimum point for the critical path in a circuit is a key task to maximize the benefit of the clock-data compensation effect.
On this section, the impact of adding 30, 100 and 190 clock stages (clock buffers) to the clock path is analyzed. Also, the same data path is going to be instantiated for each of the clock lengths.

Knowing the latencies will give a sense of the sensitivity of the path, also will help to determine the offset that needs to be added on each of the clock sources to align the clock waveform and make a fair comparison under the voltage drop scenario[5]. Table 3.2 summarize the different clock latencies for each of the experiments.

TABLE 3.2: CLOCK LATENCY FOR DIFFERENT CLOCK LENGTHS

Clock path	Clock latency
No stages	NA
30 stages	168 ps
100 stages	567 ps
190 stages	1 ns

Figure 3.14, shows how the clock period changes during the voltage drop depending on its length. On Figure 3.14(a), the voltage drop is presented, the maximum drop occurs at 57.5ns which corresponds to 100MHz frequency[6]. The voltage level at this point is ~587mV, which corresponds to ~10% voltage drop. On Figure 3.14(b), displays the clock period change on different clock trees.

It is notorious that the biggest change is given by the clock with 190 stages (clk_{190}). This clock tree stretches up to a 561ps clock period (61ps extra clock period), but in the other hand compresses about 35ps when the drop is recovering. The next stretching corresponds to clk_{100} that is about 40ps stretching and 25ps compressing. Finally, for clk_{30}, the clock tree has a very short clock path, making the clock period modulation effect weaker.

---

[5] Aligning the waveforms guarantee circuit is experiencing the same voltage level at a given time for the different clock tree lengths.
[6] This can be considered as a high-frequency voltage drop.
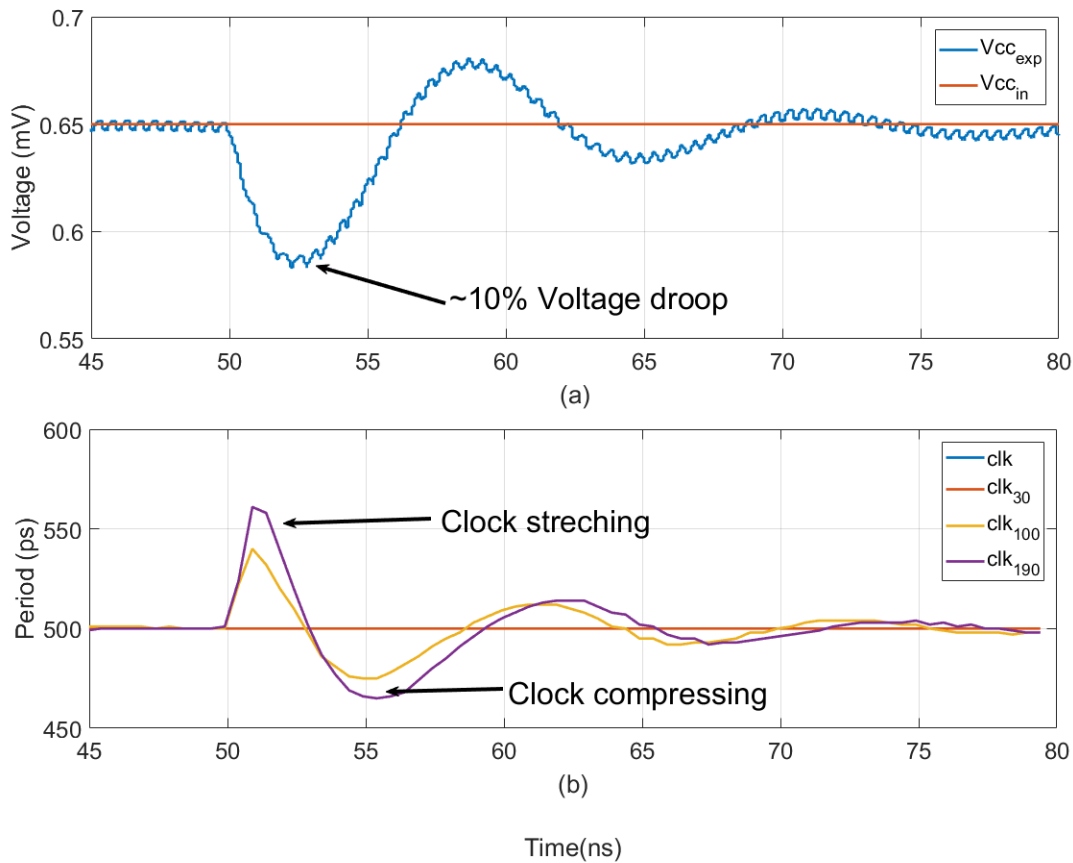
Figure 3.14: Impact of voltage drop on clock period: (a) voltage supply; (b) clock period measurement for different clock lengths.

When it comes to the slack of the path, Figure 3.15 shows how the slack changes depending on the different clock length. As it was expected, the clock of 190 stages presents the lesser slack change due to the CDC effect, dropping up to -50ps, which is greater than the most affected path driven by the clock tree (-75ps slack) without any compensation. As the clock length increase the lesser slack reduction the circuit is going to get.

Another thing to notice is the slack presents the biggest changes during the first order drop, in the subsequence drops the slacks varies but less than the first drop. For example, for the second order voltage drop, the slack in the circuit when there is not any compensation falls up to -10 ps, causing a new failure in the circuit. For the clock of 190 stages due to the compensation, the slack remains positive avoiding any failure in the circuit.
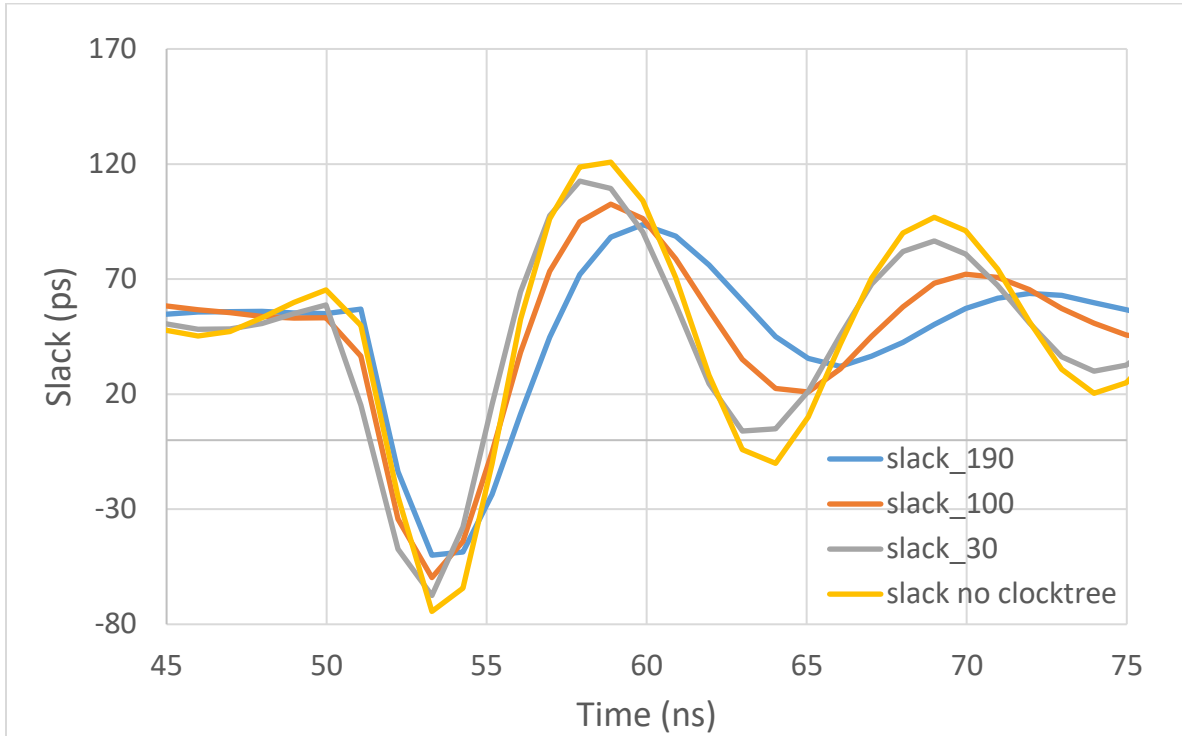
Figure 3.15: Slack impact due to voltage drop

After analyzing the slack of the data path for different clock tree lengths, the remaining item is to confirm the error signal is activated when the slack is lesser than 0ps (or even really close due to setup timing requirements). Figure 3.16 presents this data, in which the error signal for the clk_{190} path, fails until the voltage reaches its lowest value, and the rest of the cases (with smaller clock tree) the circuit fails earlier. When there is not enough compensation like in clk_{30}, the circuit tents to fail earlier with a higher voltage level, on this case is 15 mV higher.

Finally, the paths with no compensation or clk_{30}, fail even in the second order drop, which has less voltage drop compare to the first order drop.
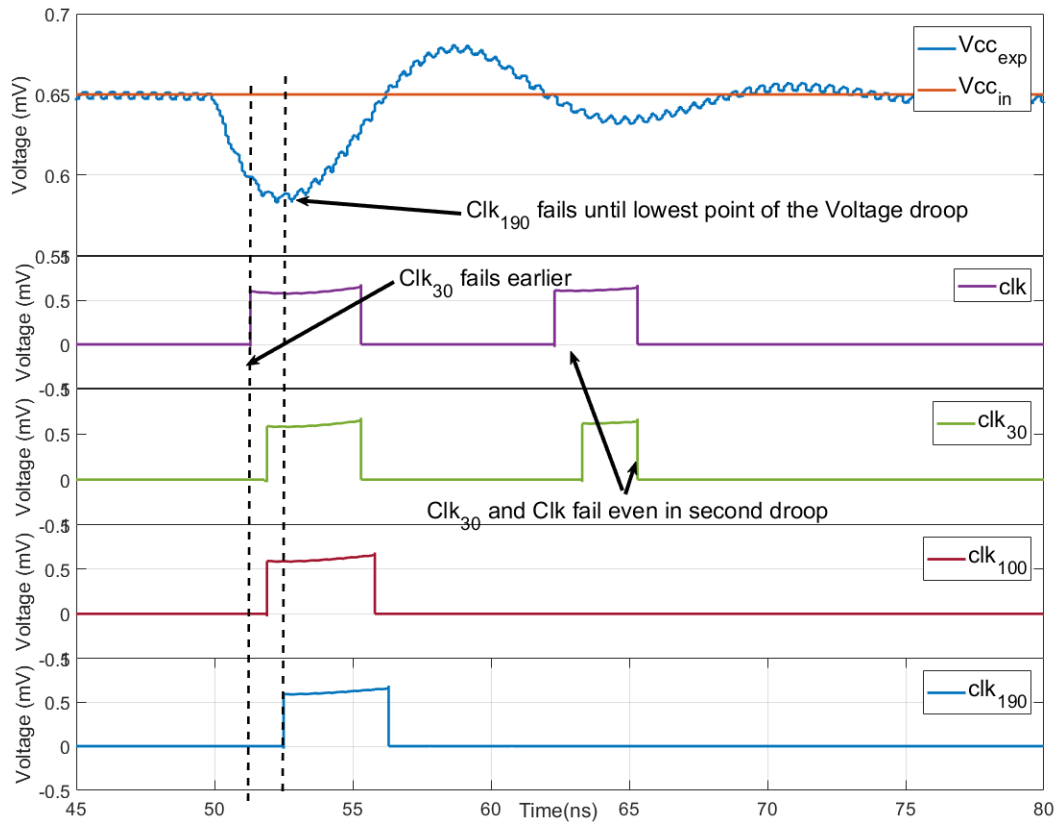
Figure 3.16: Error signal for 30, 100 and 190 clock buffers in the clock tree. Error signal occurs later when having more clock-data compensation

In this chapter, it was possible to analyze the CDC effect on the designed circuits considering a second order PDN model in a modern FinFet-CMOS technology. This effect have been analyzed through different test scenarios to establish a quantitative comparative baseline analysis, taking into account different clock tree lengths and a non-clock tree scenario. Showing a higher clock-data compensation when the clock tree sensitivity is closer to the data path sensitivity, for a very short clock path, the clock period modulation effect is weaker.

Chapter IV

Proposed Voltage Drop Mitigation Scheme

In this chapter, an adaptive and scalable technique is developed to enhance voltage drop tolerance in CMOS circuits through Adaptive Voltage Scaling (AVS), taking advantage of the clock-data compensation effect seen on Chapter III. The proposed solution is validated by applying it to different scenarios on different operating conditions with a FinFet-CMOS technology.

Figure 4.1 shows the general system to be evaluated, which incorporates the following elements:

- **Vcc_in:** power supply for the circuit coming from outside of the chip.
- **Vcc Gated:** power supply connected to the circuit/load. It is affected by load variability.
- **Load:** It is the circuit under study.
- **Variable load:** any change in the circuit translated to a decrease or increase of the current demand. Such changes can be related to turning on different block, memory access, etc…
- **Voltage comparator:** compares Vcc Gated supply with a voltage threshold (Vref). If the Vcc Gated is under this threshold, a signal will be sent to the voltage controller.
- **Voltage controller:** receives the signal from the voltage comparator and responds to it, controlling the different Power gates in the PG block.
- **Current regulation block:** It is a block that regulates the current injected to the system by varying the resistance seen between the voltage supply and the load.

The dynamic of the proposed scenario goes like this: the system possesses a load which can be certain part of a chip that can experience fluctuations in the voltage supply due to changes in the surrounding circuitry. The variable load will represent these changes through an increase or decrease of the current consumption. The voltage comparator will monitor the supply voltage level and capture these fluctuations, and based on a voltage threshold and the supply voltage level will send information to the voltage controller, which can vary the resistance seen by the power supply and inject more charge to the system through a current regulation block.
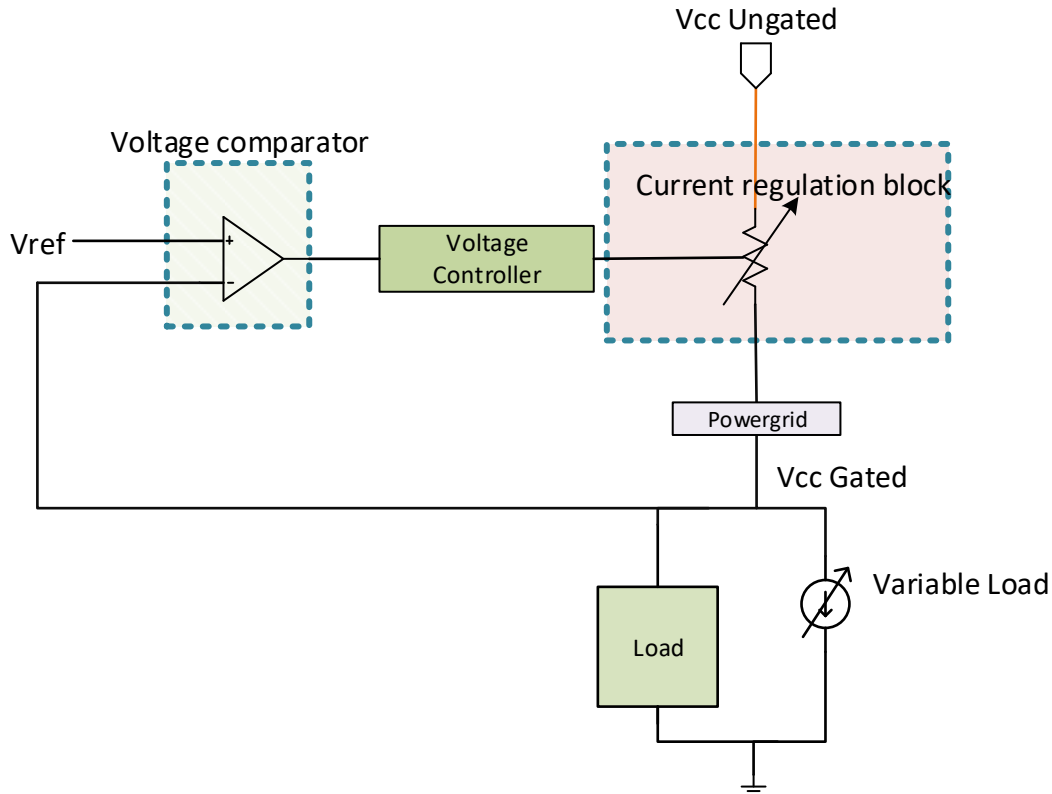
Figure 4.1: Block diagram for the voltage regulated system with a voltage comparator, voltage controller, and current regulation block in a variable load system with a powergrid model

The following section of this chapter will describe the voltage comparator design along with characterization, its response to the voltage drop and its limitations. Then, the voltage controller functionality and how interacts with the rest of the blocks will be explained; finally, the current regulation block that makes the charge injection to the system will be described.

4.1. Voltage Comparator

The voltage comparator is a block that compares Vcc supply with a voltage threshold (Vref). If Vcc is under this threshold, a signal will be sent to a voltage controller. For effects of this work, this block is called VDM (Voltage Drop Monitor).

The VDM constantly checks if the nearby[7] Vcc_Gated value crosses a programmable drop threshold, and if so, the associated drop controller turns on the local PG block to inject a local charge from Vcc_in to Vcc_Gated, and restore the voltage.

The block diagram for the voltage comparator is shown in Figure 4.2.

---

[7] Most of the voltage regulators are outside the chip. This block is really close to the circuit
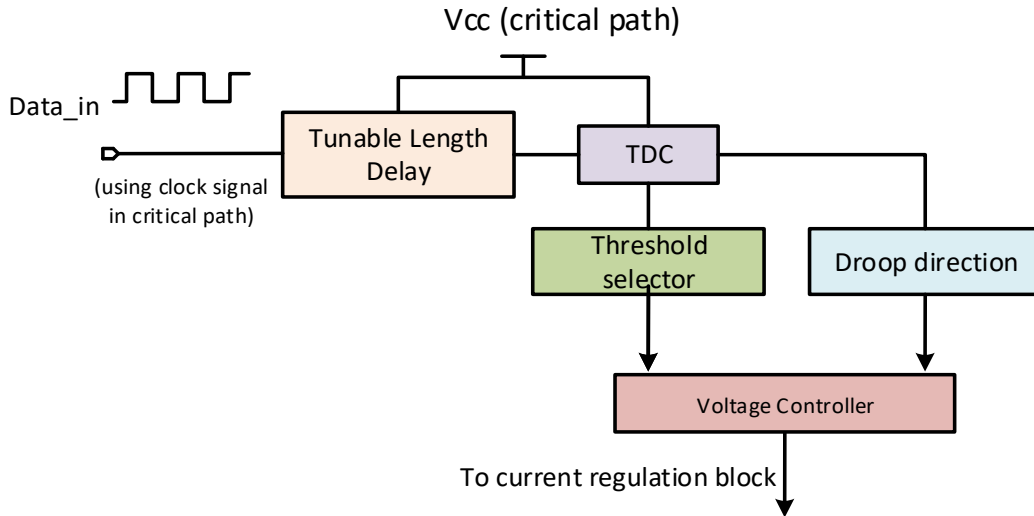
Figure 4.2: Voltage Drop Monitor (VDM) block diagram

Time to Digital Converter

The block features a time-to-digital converter (TDC). This block has been used to measure the voltage drop in the system. TDC functionality is shown in Figure 4.3, where the measurement interval defined by the start and stop signal is completely asynchronous to the reference clock signal. The measurement accuracy can be increased by a higher clock frequency. However, the higher the clock frequency the higher the power consumption for the generation and the processing of the clock signal.
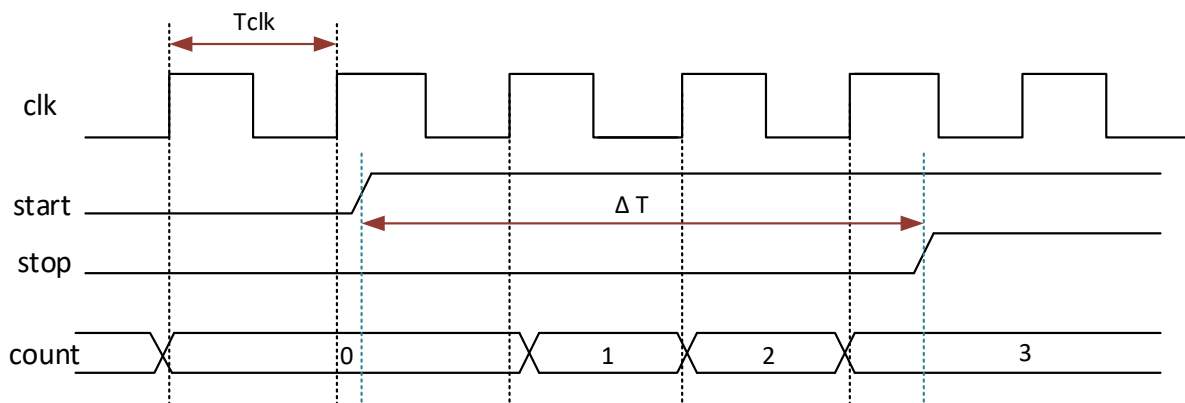
Figure 4.3: Principle of counter based TDC presenting a start/stop signals and a counter for each cycle the measurement interval is valid

Figure 4.4 illustrates the operating principle of a TDC based on a digital delay-line. A start signal is delayed along the delay-line. On the arrival of the stop signal the delayed versions of the start signal are sampled in parallel. Either latches or flip-flops can be used as sampling elements. The sampling process freezes the state of the delay-line at the instance where the stop signal occurs. This results in a thermometer code because all delay stages which have been already passed by the start signal give a HIGH value at the outputs of the sampling elements, all delay stages which have

not been passed by the start signal yet give a LOW value. The position of the HIGH-LOW transition in this thermometer code indicates how far the start signal could propagate during the time interval spanned by the start and the stop signal. Hence, this transition is a measure for the time interval.
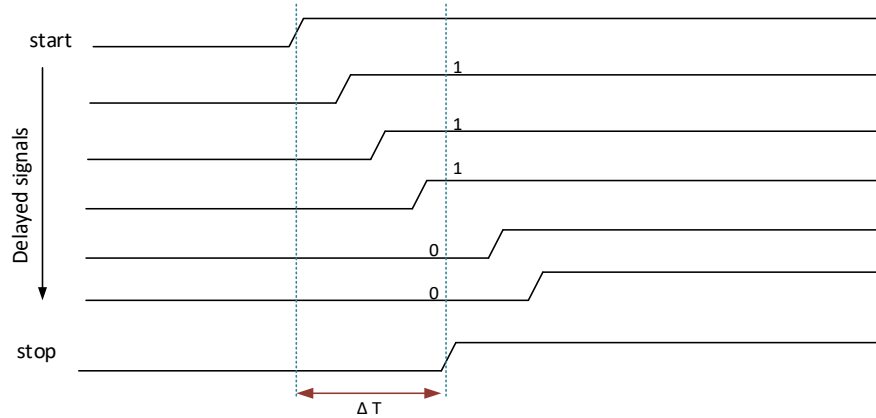
Figure 4.4: Principle of a time-to-digital converter using delayed versions of the start signal

An implementation of the basic delay-line TDC is shown in Figure 4.5. The start signal ripples along a buffer chain that produces the delayed signals. Latches are connected to the outputs of the delay elements and sample the state of the delay line on the rising edge of the stop signal. The stop signal drives a high number of latches so a buffer-tree (not shown) is required. Any skew in this buffer-tree directly contributes to the non-linearity of the TDC characteristics. More details about the clock tree in the TDC are found in Appendix D. The delay elements for the delay line in this case are buffers. Other type of gates like inverters can be used, however, this will impact on the rise and fall times for the signal traveling across the delay line, creating an unbalance. Not having a balanced delay for both rise and fall can cause wrong behavior of the code, in which the code will not be a straight pattern of 0's and 1's. More details about this can be found in Appendix A.
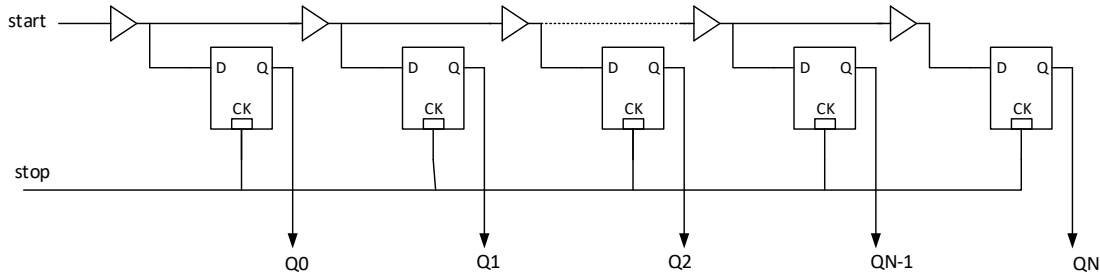

Figure 4.5: Implementation of a basic delay-line based time-to-digital converter using buffers and latches as sampling elements

If we use the clock signal as start and stop signals, plus the fact latches are being used as sampling elements, the total time to measure the TDC (ΔT) will be:

$$\Delta T = \frac{T_{clk}}{2} \tag{4.1}$$

This will give half period[8] to convert any delay change into a digital code. Hence, in a clock cycle the TDC will do the delay measurement on the first half (high face of the clock), and for the second half the code is read and store in a sampling element.

TDC measurement   TDC code read
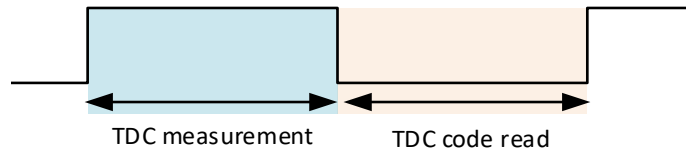
Figure 4.6: TDC functionality during the clock cycle when TDC measurement is done when clock is high and the code read is done when clock is low

Tunable Length Delay block

It performs an "offset" for the TDC, in case we need to overcome any timing issues between the clock and the TDC at the time to test the circuit. However, the delay to measure cannot be greater than the offset plus the resolution of the TDC, in such case we need either to decrease the offset seeing or are more stages to the code of the VDM (adding more gates to it). This circuit goes as in Figure 4.7.
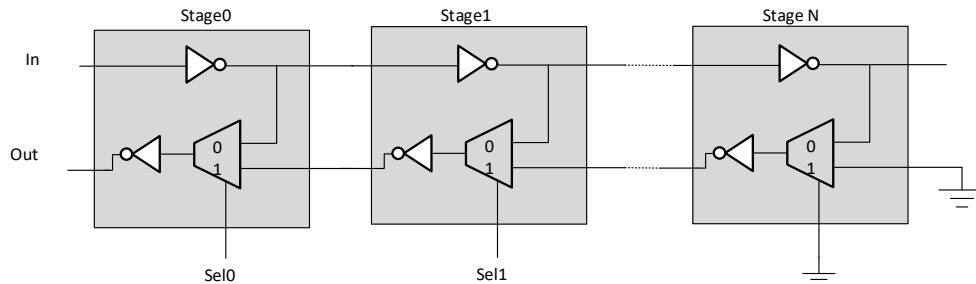
Figure 4.7: Tunable Length Delay block with multiple stages enabled by a selection signal

With this kind of topology, the number of stages can be added in a modular way in order to increase the delay in the circuit. Per stage there are two units of delay (delay of an inverter) plus the delay of the mux. The last stage will be "hardcoded" to make the return path.

4.1.1. VDM Characterization

The VDM characteristics are presented in Table 4.1:

TABLE 4.1: VDM SPECIFICATIONS

Parameter	Value
Tunable length delay stages	16 stages (set by Sel[14:0] signal, last stage is hardcoded)
TDC resolution	64 bits

---

[8] Latch is transparent during this period

The offset value cannot be greater than half code cycle[9]. In this case, Sel is equal to *15'h000f*, the VDM code can be read in an accurate way. Using this value, the output thermometer code looks like shown in Fig. 4.8.
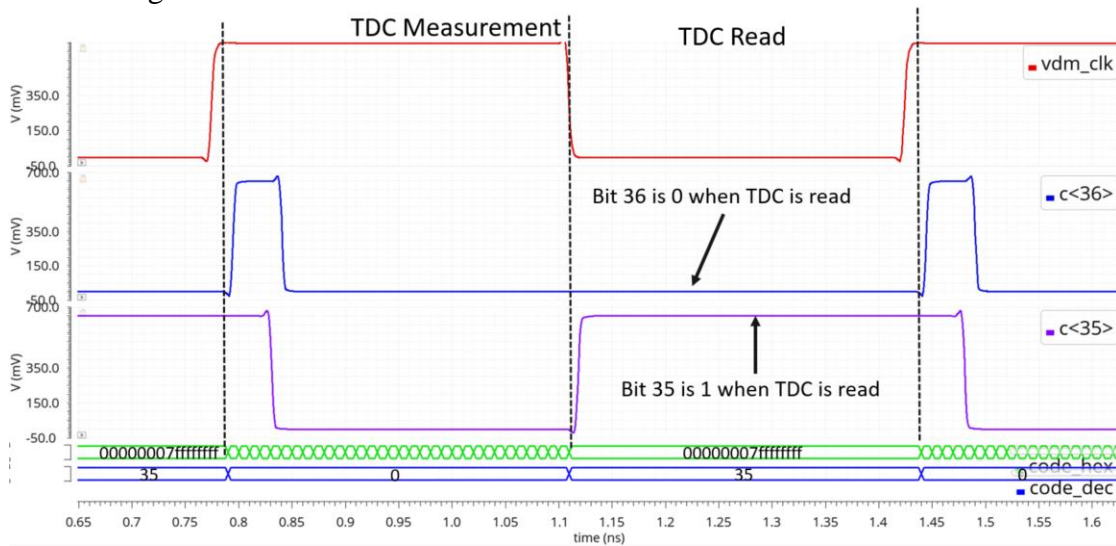
Figure 4.8: Output code when supply voltage is 0.585 V at1.5GHz with offset 15'h000f

In Figure 4.8, the output code is *64'h00000007ffffffff* (35 in thermometer code). It is observed how output 35 is ON when the TDC is reading, but c<36> is OFF. Now, the VDM is going to be characterized at 1.5 GHz[10]. This will give a perspective what is the expected code. In Table 4. 2, the different codes depending on the input voltage. Highlighted are the values of interested; also, a 1's counter translation is done in order to make the code more readable.

TABLE 4. 2: THERMOMETER CODE FOR THE VDM @1.5GHZ

| Voltage (mV)	VDM Thermometer code	1's counter translation
550	0x0000 0000 07ff ffff	27
560	0x0000 0000 1fff ffff	29
570	0x0000 0000 7fff ffff	31
580	0x0000 0003 ffff ffff	34
585	0x0000 0007 ffff ffff	35
590	0x0000 001f ffff ffff	37
600	0x0000 003f ffff ffff	38
610	0x0000 01ff ffff ffff	41
620	0x0000 07ff ffff ffff	43
630	0x0000 3fff ffff ffff	46
640	0x0000 ffff ffff ffff	48
650	0x0003 ffff ffff ffff	50
660	0x001f ffff ffff ffff	53
670	0x007f ffff ffff ffff	55
680	0x01ff ffff ffff ffff	57
690	0x07ff ffff ffff ffff	59
700	0x1fff ffff ffff ffff	61

---

[9] If it is greater than this the VDM won't give accurate results

[10] In case other operating frequency need to be used, VDM needs to be re-characterize

Having code 50 when the system is stable, and 35 when the system has reach the ~10% drop from the initial supply value. With this kind of monitor, the system presents lot of granularity for voltage changes. This allow having multiple thresholds. Therefore, the system can respond in different ways depending of the voltage level. This also gives an advantage versus the operational amplifier, which has only one threshold value.

4.1.2. VDM Response to the Voltage Drop

The VDM is connected to the Vcc domain that is experiencing the voltage drop and is giving an output code that changes every clock cycle depending on the voltage level. The VDM will be connected as is shown in Figure 4.9.
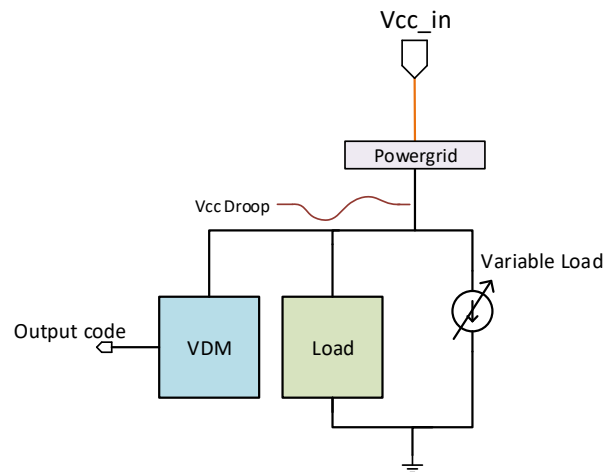
Figure 4.9: Scheme of VDM detecting voltage drop in a variable load system with a powergrid model.

Figure 4.10 shows VDM code is changing dynamically depending on the voltage value. However, this voltage value is not fixed at the "TDC measurement" stage, this voltage is changing constantly. Hence, the output code will differ a little from the code shown in Table 4. 2, where the voltage is fixed at a certain value. The code will be stable for the low face of the clock and in the high face is when all the calculation is happening through the TDC.
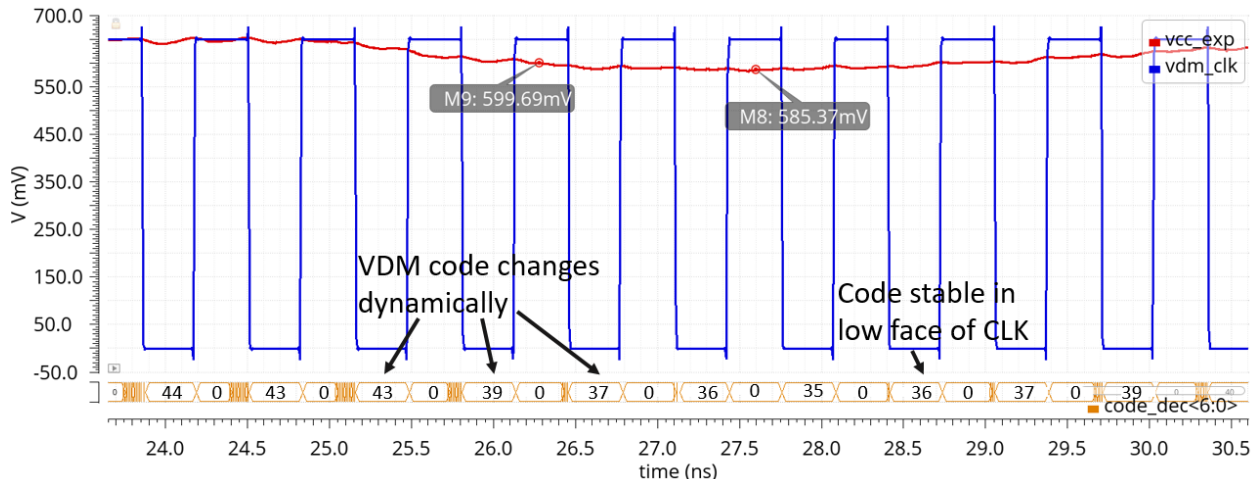
Figure 4.10: VDM output code changes dynamically based on the voltage level doing outputting a stable code when clock signal is low

At 27.5ns, the voltage level is 585mV and the VDM code is 35, which matches with the results in Table 4.2. But, at 26.5 ns, it is shown how the VDM code is 37 for a value around 600mV, and this not exactly the same as in Table 4.2, due to the dynamic variations of the voltage.

As It was explained in Chapter 2, high frequency voltage drop can be between 50-300 MHz. In this case frequency of the voltage drop is 100 mHz. However, in Appendix C is shown the response of the VDM under other frequencies.

4.1.3. Limitations and Considerations of the VDM

For the "data_in" signal of the VDM is needed a toggling signal. If the clock signal from the circuit that is experiencing the voltage drop is used as the same clock signal of the VDM, this causes any clock-data compensation is happening on the system will also affect the code reading in the TDC clock. This is because on the first half of the drop there is a stretching of the clock period, and then a compressed waveform in the second half when the drop is recovering. This gives a different ΔT for the time-to-digital conversion, causing an asymmetric code during the voltage drop event. Hence, we need to make sure the window in which the latch is transparent, it is constant across all sample point where the voltage drop is occurring.

In Figure 4.11, asymmetric codes for the VDM are shown, where the same voltage is having two different codes depending on the direction of the voltage drop. For the first half, the code is greater than the second half codes. This is because the second half presents a compressed clock period for the digital conversion, while the first half has a stretching in the clock period. For example, for 600mV at 26ns, the code is 43 when the voltage is dropping, but when it is recovering (compressed clock cycle) the code is 35, which is by the way the same code as in Figure 4.10 when the voltage is 585mV. Another example is at 29ns when the code is 34 (for 591mV) that is even smaller than the code when the voltage has its lowest value (585mV).

For example, for 600mV at 26ns, the code is 43 when the voltage is dropping, but when it is recovering (compressed clock cycle) the code is 35, which is by the way the same code as in Figure 4.10 when the voltage is 585mV. Another example is at 29ns when the code is 34 (for 591mV) that is even smaller than the code when the voltage has its lowest value (585mV).
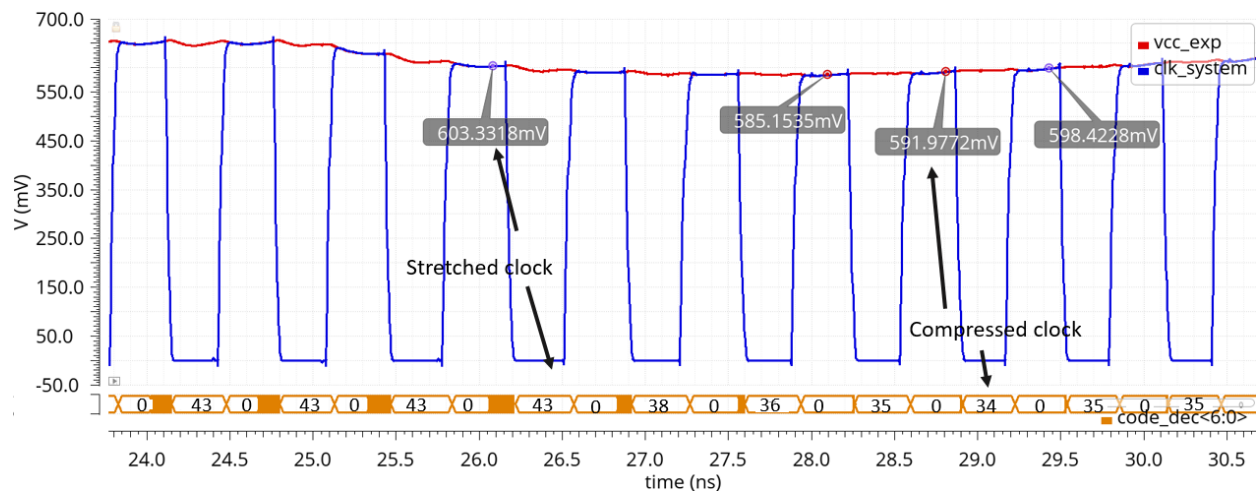

Figure 4.11: Asymmetric codes for VDM due to CDC effect having an expanded clock period when voltage is dropping and compressed clock when voltage is recovering

Other limitation of the VDM is where the magnitude of the voltage drop starts to increase beyond the initial spec of the 10%. In Figure 4.12, it can be observed how the VDM starts presenting aliasing, giving a wrong code when voltage drops less than 565 mV (around 14% voltage drop). A wrong code in this case means not having a straight pattern of 0's next to a straight pattern of 1's. If the code starts being at intercalated patterns of 0s and 1s, this will be considered as an error. Based on Figure 4.4, the output code is a shift register across all the latches in the VDM. When voltage starts getting to lesser than 505 mV, the bits in the VDM code start to overlap.

Starting at 6.1ns the encoder is stuck at 0, which means the output code for the VDM is not having an expected value.
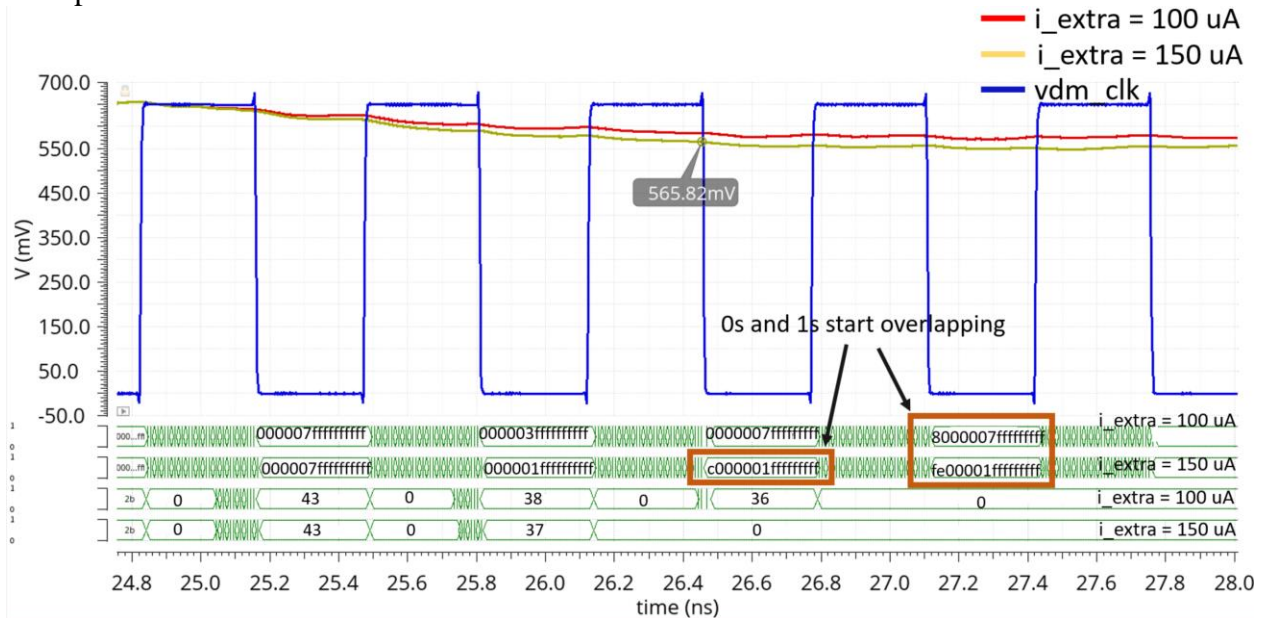
Figure 4.12: Voltage drop sweeping for the VDM

In Figure 4.12 for the hexadecimal codes, it is observed how the MSBs in the code for the 14% voltage drop start overlapping with the LSBs in the code. The voltage level is so low that causes the transition from 1 to 0 not been propagated throughout all the bits in the code, which is not the same case when voltage is greater than 565 mV.

In order to avoid aliasing, one possible solution is to relax the frequency of the VDM to give enough time to the previous pulse to travel across all the registers in the VDM. However, the trade off with the solution is the reaction time of the controller. Depending on the operating frequency of the circuit, the new frequency of the VDM might be too slow to react before the critical path in the load fails. The objective with this scheme is to have sampling frequency equal or faster than the operating frequency of the load, otherwise the correct behavior is not guarantee. Another possible solution can be to reduce the number of register in the VDM. At the time to have a shorter register chain[11] will avoid this overlapping in the MSBs in the code,

---

[11] In this case 64 registers were used, so a number lesser than this will avoid the aliasing

4.2. Voltage controller

The voltage controller manage the current injection based on a user defined threshold and the VDM code and its previous value. It can determine if the system is experiencing a voltage drop or not. Since both values present and past of the voltage are an input to the controller, the voltage controller can determine if the drop is going down or up.

When the voltage exceed the threshold and its going down, the voltage controller will inject more current to the system through the current injection block. In the other hand, if the voltage is exceeding the threshold but is going up, this means the voltage is recovering then in order to save power will turn off the current injection block. This is also explained in the diagram presented in Figure 4.13.
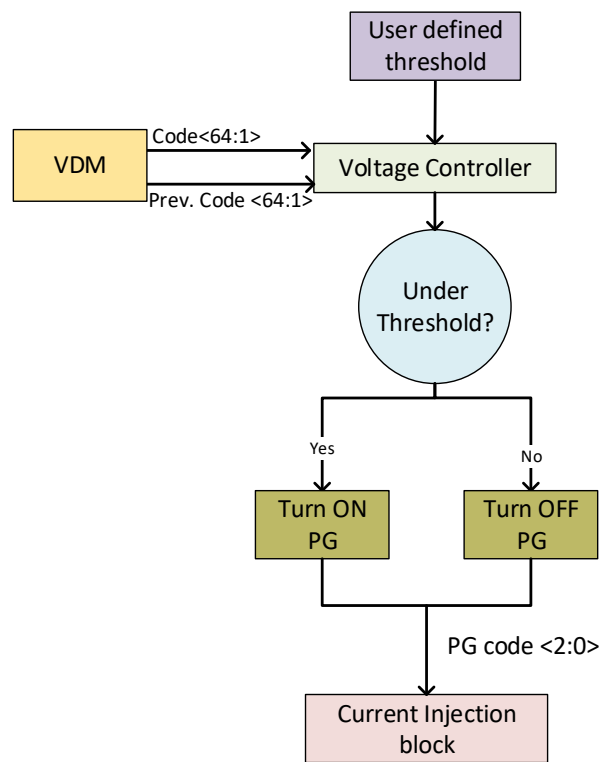
Figure 4.13: Voltage controller flow diagram when a voltage drop event is happening.

In order to determine this threshold, we need to determine how much extra current or voltage drop the system can tolerate. Based on Figure 4.14, the system can tolerate only 100µA extra current without failing. This current step translates to a voltage drop of 7.7% (~600mV).
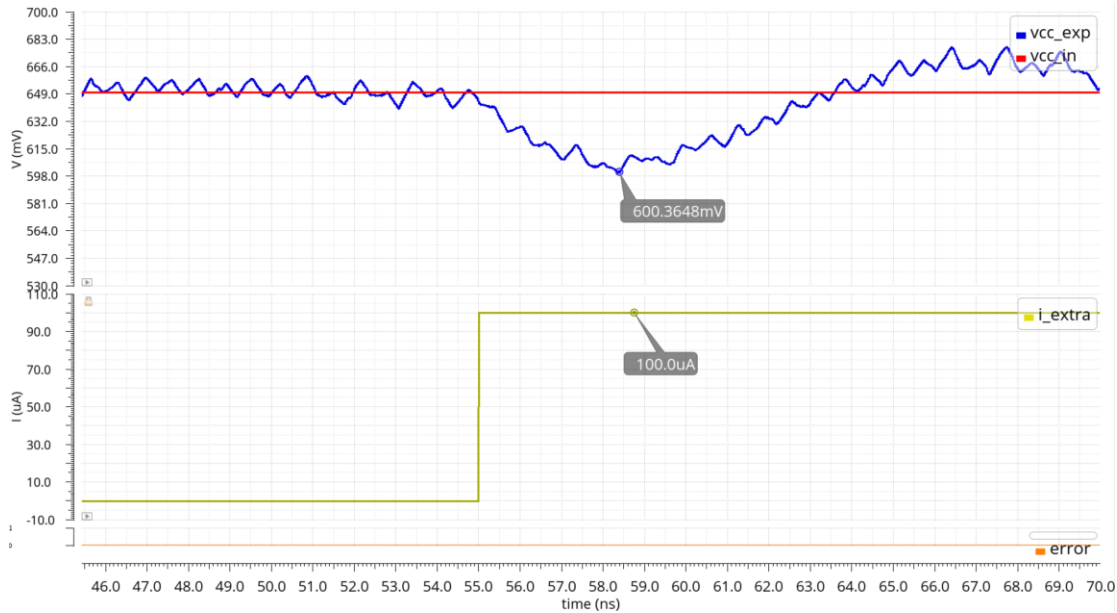
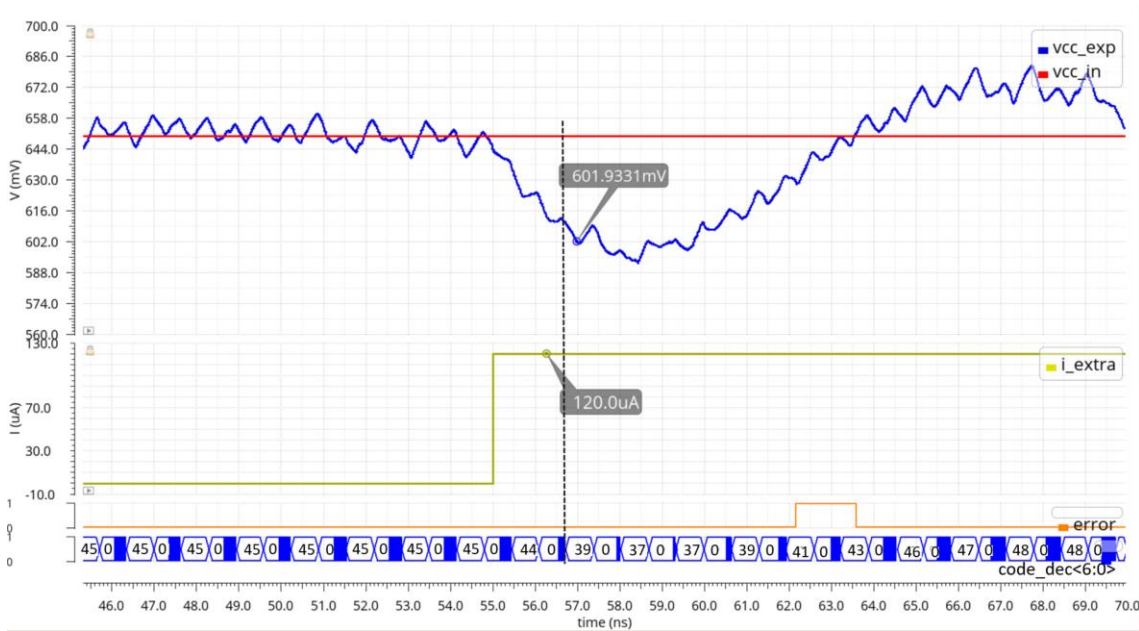**Figure 4.14:** Maximum tolerance by the system when there is no voltage compensation using a voltage controller and a current injection through power gates

Applying a higher current step such like 120µA (Figure 4.15), the system will fail by activating the error signal.

Figure 4.15: Minimum current steps that cause failures when there is no voltage compensation using a voltage controller and a current injection through power gates

Since the maximum voltage drop allowed by the system is 600mV, this translates to "39" in the VDM code, which means the voltage tolerance scheme needs to be activated before this by setting a higher code, lets say 41. Having a higher code means the voltage controller will react right before its maximum threshold. The controller will turn on PGs to inject more current once the voltage threshold is exceeded. In Figure 4.16, it is observed how the "pg_enable" signal activates once the

voltage is lower than 600mV at the time to have a 120mV. The error signal stays at 0, meaning there is no error at the time to have the current event on the system.
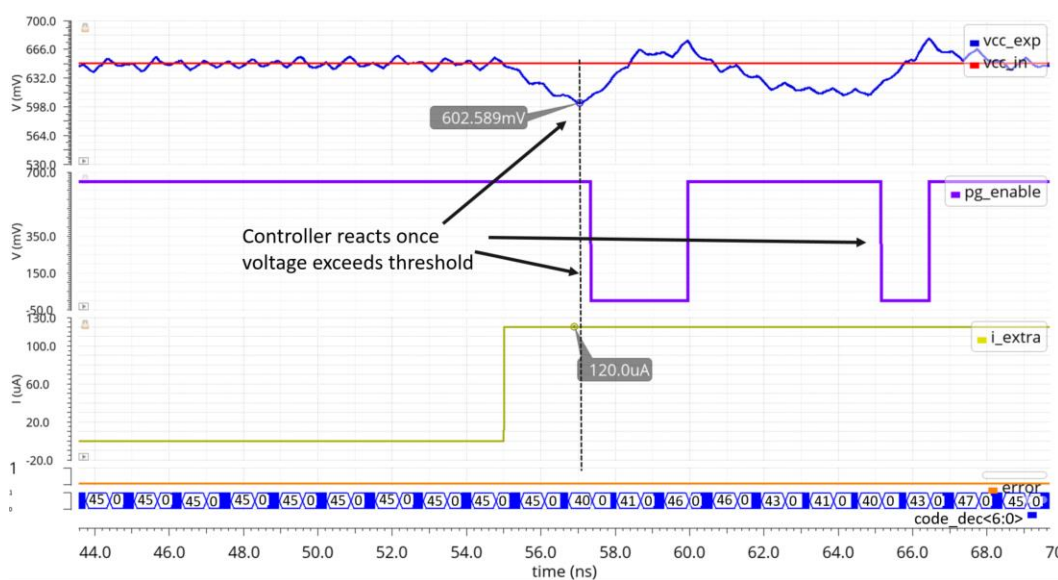
Figure 4.16: Circuit response under a 120µA current step when using a voltage controller and a current injection through power gates

4.3. Current Regulation Block

PMOS can be used to inject charge to the system [55] in this configuration they are called power-gates (PG). The PG will give more flexibility to control the design. Basically, they can be sized in such a way the voltage drop can be controlled depending on the number of PMOS are turned ON or OFF. Now, this PG block can be placed in several locations. It can be between the powergrid and the clean power supply (Vcc In), also between the powergrid and the load, and the final configuration will be having an additional power supply which will be clean and it can have same or higher level as "Vcc In". This power supply will have its isolated powergrid with different parasitics than the regular powergrid. The system for the three models is shown in Figure 4.17.
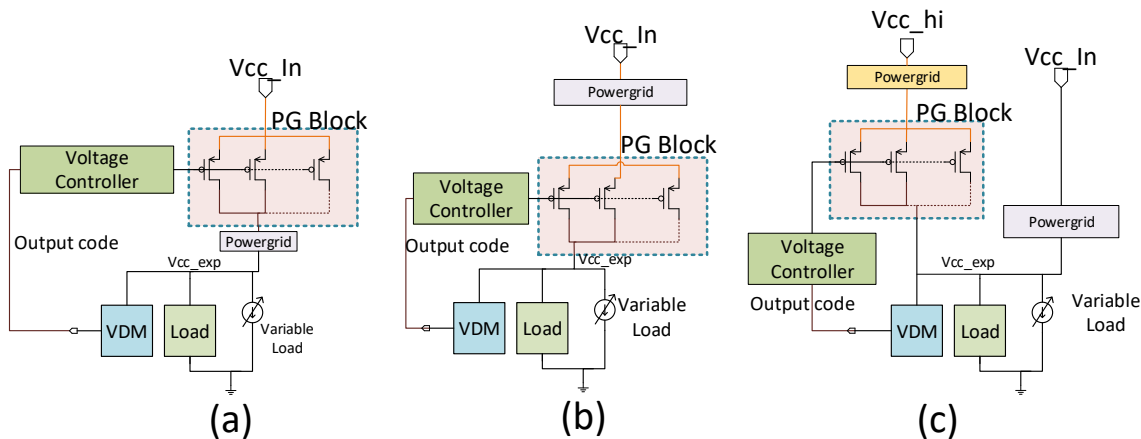
Figure 4.17: System with PG Block and VDM. (a) PG between the powergrid and Vcc_In (b) PG between the powergrid and load. (c) PG connected to the load and the Vcc_hi power supply

Configurations in Figure 4.17(a) and Figure 4.17(b) present a problem since the PG work as a Low-Drop Out (LDO) voltage regulator. It doesn't matter how many PGs are turned on, the only thing that will happen is the "current injection" to the system. Adding more PGs will reduce the resistance that Vcc_In is seeing to Vcc_exp domain, but that doesn't mean Vcc_exp with have a different behavior than the node connected to the powergrid. In the best case scenario Vcc_exp will be equal to Vcc_In.

If an array of PGs are connected to the system in which some of them are turned on by default, and also there are more PG that will be activated by an enable signal the response of the system will be like in Figure 4.18. It is observed how the Vcc_exp follows Vcc_pg with an offset. This offset is the voltage delta across the PGs.
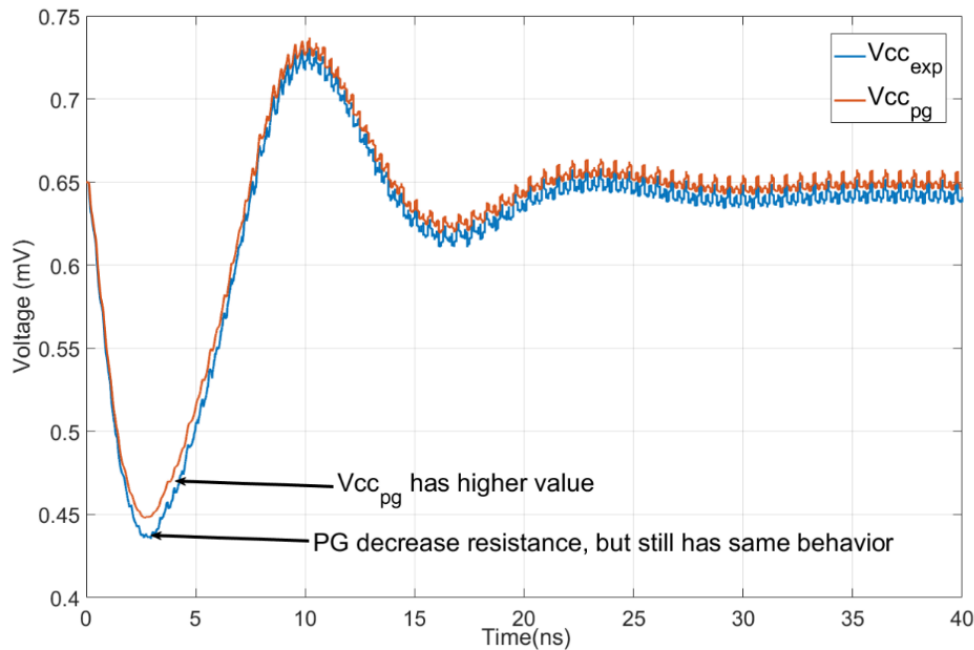
Figure 4.18: Voltage drop for scheme a and b of the power gate block when the current injection is not done from a clean power supply

If the configuration switches to the scheme presented in Figure 4.17(c), now the charge injection is being made directly from the Vcc_hi clean supply, hence is not being affected by the voltage drop. This modifies the behavior of the voltage supply in the system. Depending on how many power gates are turned on, the system will respond differently, it can either respond fast, slow or even can't be enough to compensate the voltage drop. Turning on more PGs will decrease the resistance for the injected current when the voltage is lower than the voltage threshold.[12]

---

[12] Which is 585 mV in this case.

(a)

(b)

(c)
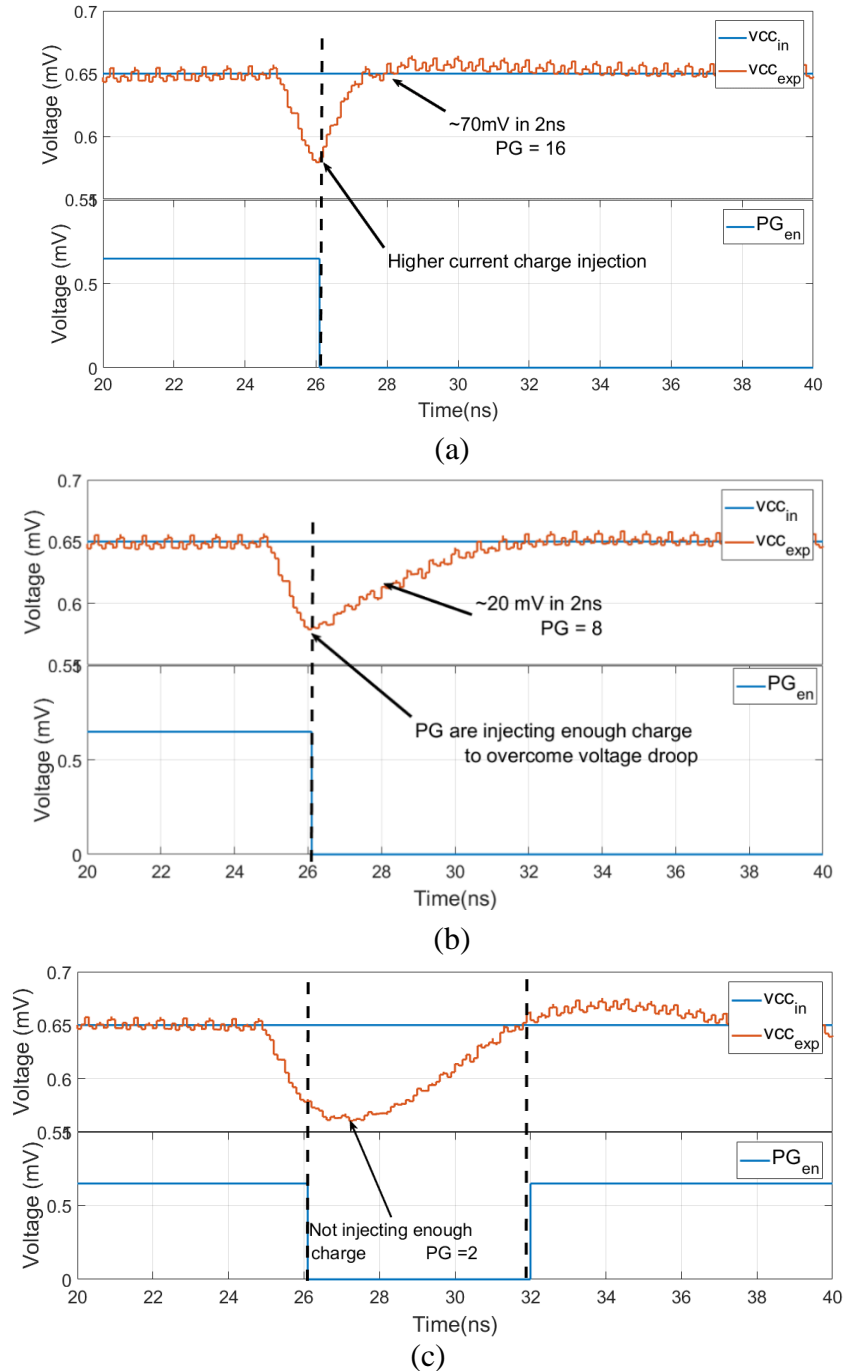
Figure 4.19: System response depending on the turned on PGs: (a) Fast Response; (b) Slow response; (c) Not enough current injection

Figure 4.20 shows how PG size can be swept in order to find the minimum size to avoid any overdesign in mitigation scheme. Having more PGs than necessary translates to extra area and leakage power consumption. The minimum PG number to inject enough charge to the system in order to star recovering the voltage is six.
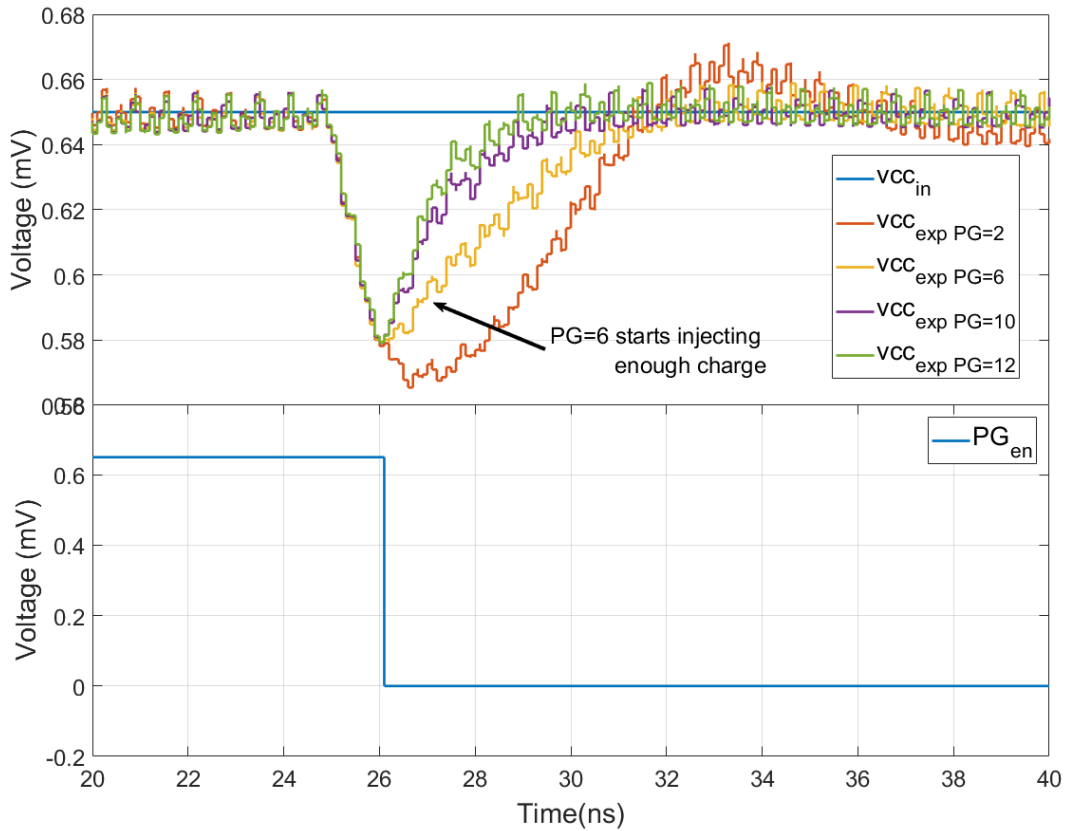
Figure 4.20: PG size sweep for the current regulation block optimizing the minimum numbers of PGs needed to start injecting enough charge to the circuit to overcome voltage drop

The VDM creates a code depending on the voltage value by doing a time-to-digital conversion. The circuit allows having a multi-threshold operation plus voltage drop direction detection, in which the controller can react depending on the voltage level. Having that low voltages can cause a misread of the voltage code, causing incorrect behaviors of the voltage controller. This voltage limitation on the controller depends on its frequency. The VDM will be implemented in layout, and different case of study are going to be evaluated through post-layout simulations on the next section.

Chapter V

Voltage Drop Tolerance Circuit Testing

In this chapter, a layout implementation of the circuits designed on Chapter 3 and 4 is created, followed by parasitic extraction and post-layout simulations. This will present more realistic results, and will create a model of the system closer to reality. In order to do this, several data paths are going to be tested to understand the benefits and limitations of the proposed technique.

5.1.   Simulation Framework Overview

One of the stages at the time to design a micro-processor is to perform Static Timing Analysis (STA). STA is all about analyzing the cell delays and net delays over millions of paths in a design and fixing if any violation arises in those paths by comparing with the timing constraints. These paths possess a slack based on the timing constraints and the delay of the path from the start point to the endpoint. Having said that, there can be multiple paths with different slacks, one of the reasons for this can be the different delay across the timing path or different clock skew.

In a design, there can be found slow paths (which are usually called the critical path), paths with medium delays, and fast paths. Figure 5.1 shows a high level view of the different timing paths in a design.
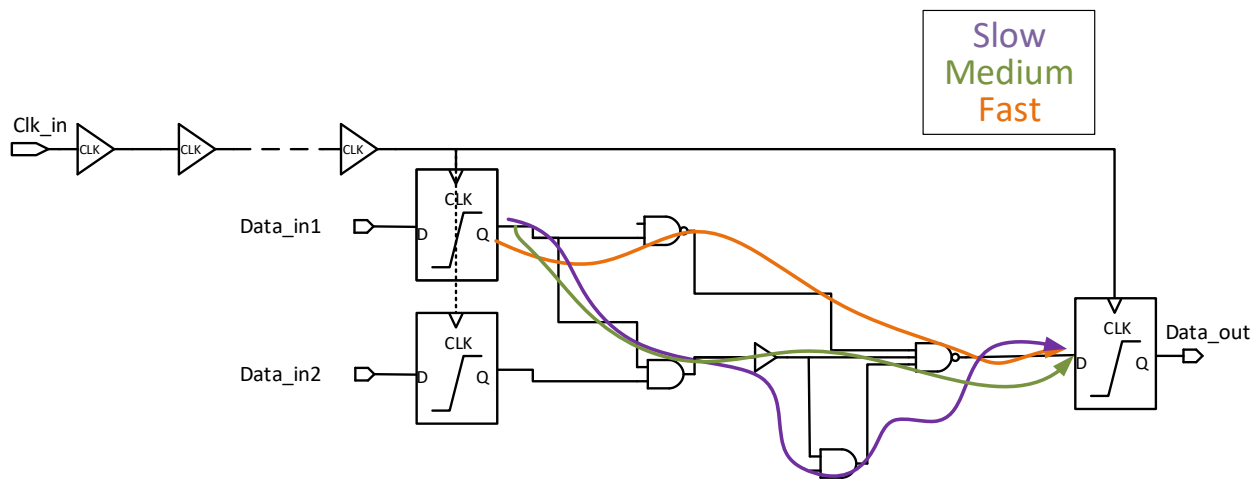
Figure 5.1: Different path delays across the same design depicting a fast, medium and slow timing path when performing STA analysis

In order to present slack variety in the design, the data path was modified in such a way that can be tunable at the time to do testing without modifying the layout, which causes an extra overhead. Having different delays in the data path will create different scenarios, making possible the analysis of the voltage drop tolerance scheme at the time to have different timing slack.

The tunable data path possess seven different stages. Each stage of the tunable data path possess a selection of different number of buffers. Also, depending on the type of stage (gross, medium or fine) the delay of the added buffers can be higher or smaller.

TABLE 5.1: TUNABLE LENGTH DATA PATH DESCRIPTION

Stage	Description
Offset	30 buffers
Gross Tunable stage 1	Selection of 11-8 buffer stages, with a mix of P-stacked high delay buffers and regular buffers.
Gross Tunable stage 2	Selection of 8-5 buffer stages, with a mix of P-stacked high delay buffers and regular buffers.
Medium Tunable stage 1	Selection of 6-4 buffer stages, with a mix of P-stacked buffers and regular buffers.
Medium Tunable stage 2	Selection of 4-2 buffer stages, with a mix of P-stacked buffers and regular buffers.
Fine Tunable stage 1	Selection of 2-1 regular buffer stages
Fine Tunable stage 2	Selection of 1-0 stages
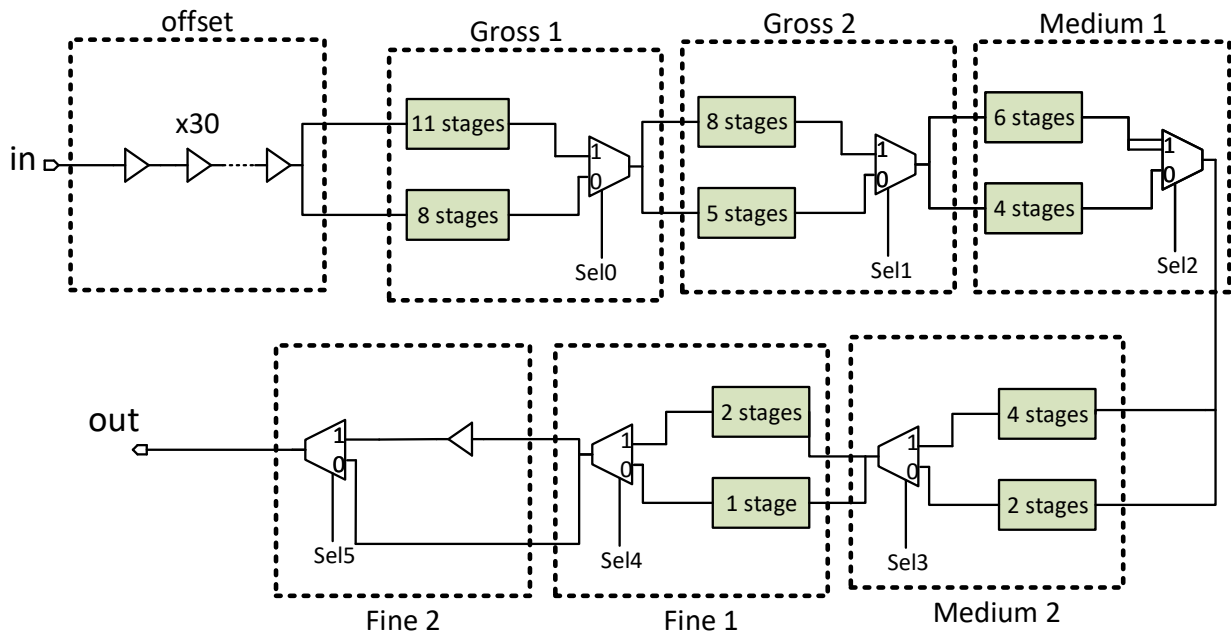
Figure 5.2: Tunable length data path with multi-stages for gross, medium, and fine-tuning.

With this, the design has the flexibility to test the mitigation scheme under different conditions in which the data path possess different timing margin. The maximum delay of the data path is when all stages are activated (62 stages) and it is about 655ps, and the minimum delay is 520ps (50 stages).
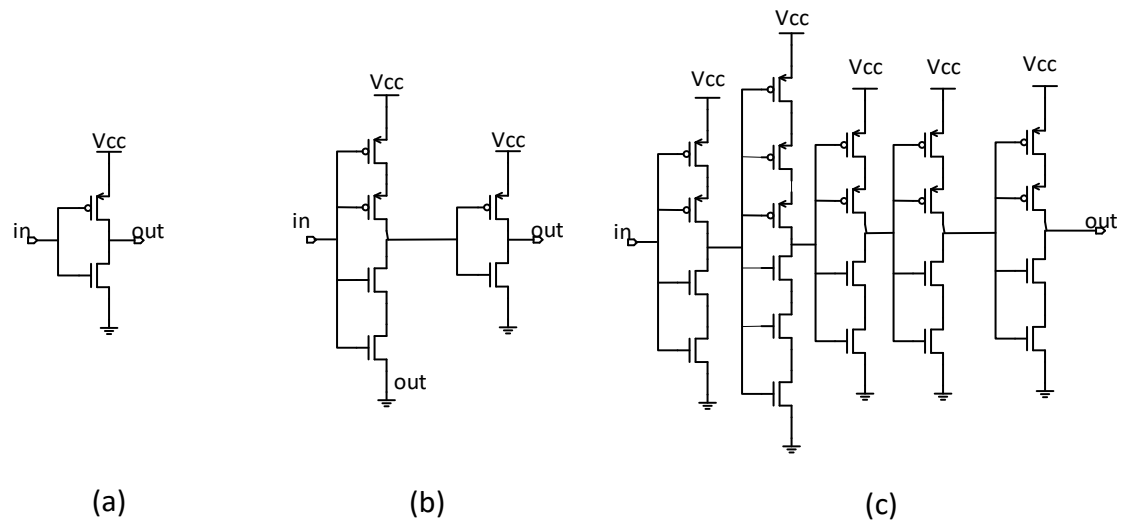
Figure 5.3: Different types of buffers used in the tunable data path: (a) regular buffers; (b) p-stacked buffers; (c) p-stacked high delay buffers.

The device under test (DUT) is composed by the tunable data path in which can select multiple delays depending on the "sel_delay" signal, a pre-defined the clock tree[13] (that maximizes the CDC effect and helps to give more time to the voltage controller to react), and finally the error detection signal, which is activated in the low face of the clock. In Figure 5.4, a high-level schematic for this is illustrated.
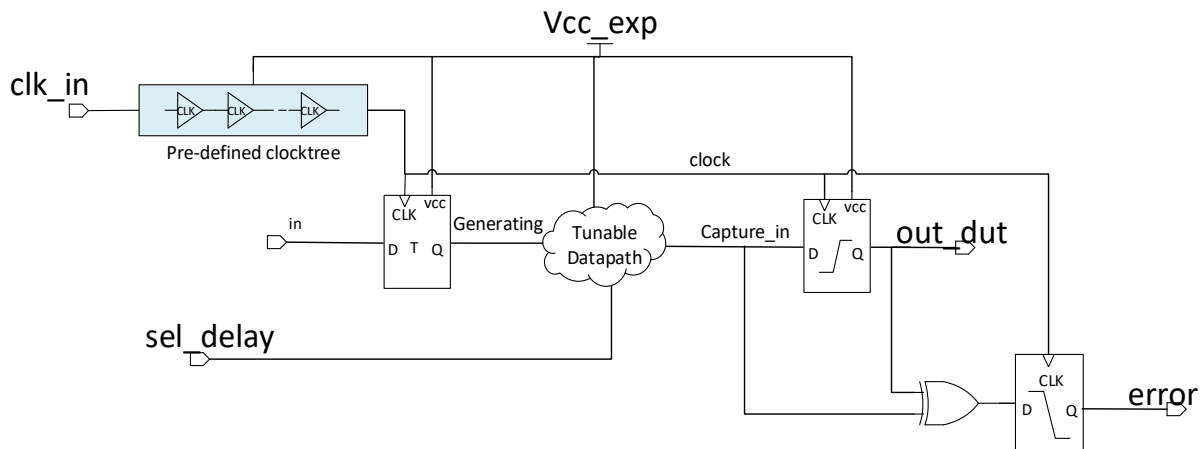
Figure 5.4: High level schematic for DUT built with a clock tree, launching and receiving flip-flops, a tunable data path and error signal generation
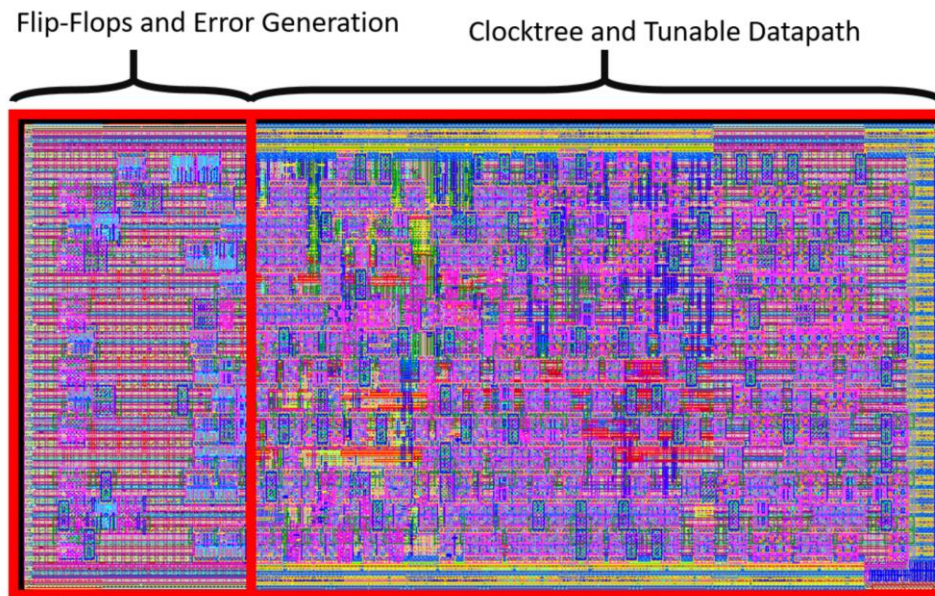
5.2. Layout Implementation

The layout implementation for the tunable data path and the voltage drop tolerance circuit are presented on this section. Both are created separately, and each of them have their own constraints. The layouts are created using ICC2 Synopsys® tool.

---

[13] The clock tree has 190 stages based on the studies done in Chapter 3

Tunable Data path

There is no optimization on this layout more than a few custom placement cells. This because we wanted to replicate a critical path in which there was no floorplan or routing optimization[14]. This will add a certain variation to the design, creating a mix of device and interconnect delay across the circuit. This layout is presented in Figure 5.5.

The receiving and launching flip-flops are located really close to each to avoid any clock skew, same with the flip-flop for the error signal generation. The rest of the tunable data path is on the right side. The placement and routing was done by the tool without any timing constraint. The area of the circuit is around 72 μm^2.

Figure 5.5: Tunable data path layout implementation on ICC2 Synopsys® (ICWEB view)

<u>Voltage Drop Mitigation Scheme</u>

Compared to the tunable data path, this circuit possess a few portions of the design that are timing critical. One of them is the TDC, this block needs to have a very regular placement and routing for the delay line cells and the receiving latches (refer to Figure 4.5). They need to be delay matched, otherwise, at the time to have a delay mismatch, this will lead to a wrong code for the voltage controller. Having an erroneous code will prevent a good functionality of the voltage controller. The layout for the Voltage Drop Mitigation scheme is presented in Figure 5.6

---

[14] This mimics the worst case scenario
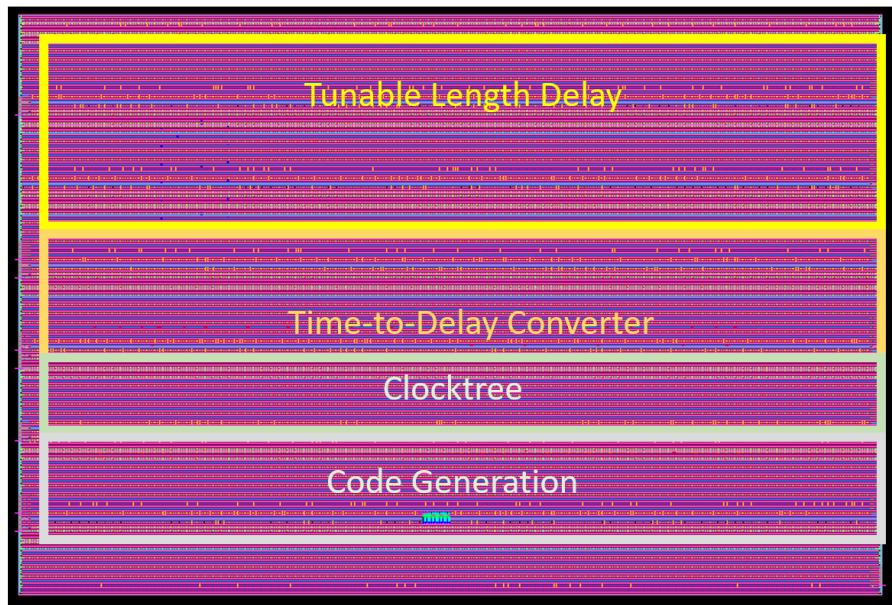
Figure 5.6: Voltage Drop Monitor Scheme Layout implementation on ICC2 Synopsys®
(ICWEB view)

The whole design has a very regular placement, but there is a special focus on the TDC portion that is a key block for the time-to-digital conversion, and also feeds the voltage controller with the code that corresponds to the voltage level. The area of the circuit is around 390 µm^2.

In Figure 5.7, it is observed how the placement is done with a uniform pattern, and also how the dataflow goes very systematic across all cells. Same with routing, which is done very similar across all the connections. This will prevent to not having a correct code for the voltage controller. The buffers of the delay line go in diagonal direction up and down. Also, they are aligned with the receiving latch, facilitating the tool to do a uniform routing across the whole module.
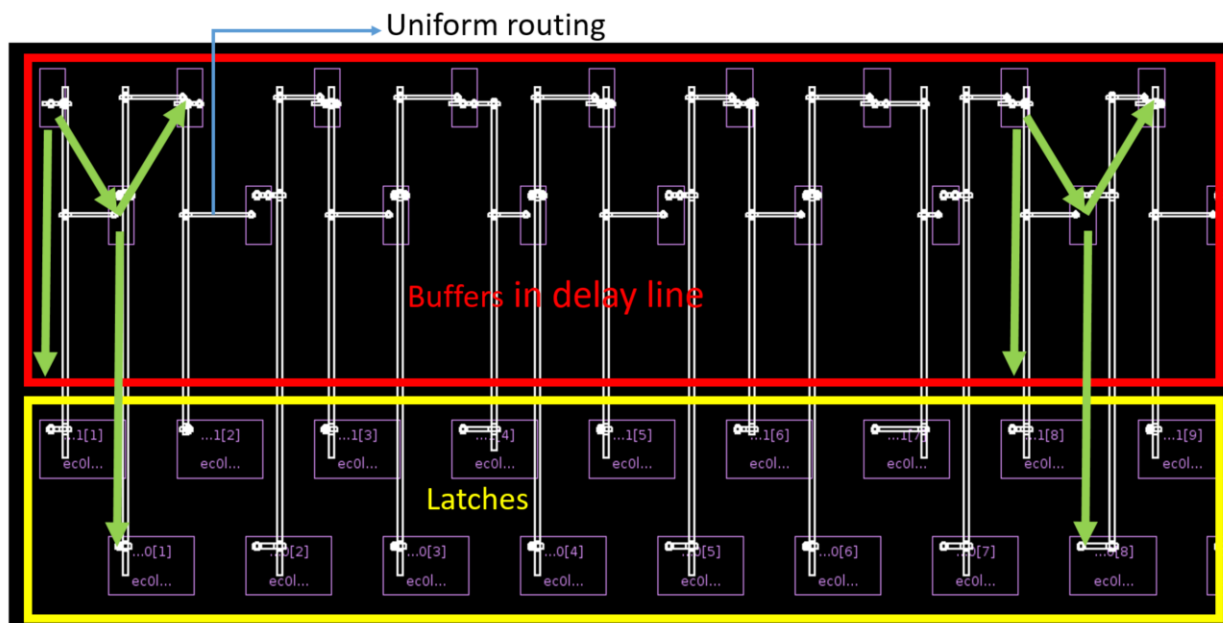
Figure 5.7: TDC custom placement delay line buffers and latches showing a regular placement and routing among all the stages.

After implementing the VDM in layout, there were some modification that needed to be done like sizing of the clock tree for the latches, and extra buffer insertion to solve hold issues in the flip-flops that outputs the VDM code. More details can be found in Appendix B and D.

5.2.   Post-Layout Simulations Setup (POLO)

The parasitic capacitances and resistances extracted might be critical in affecting the actual performance of the design. Performing a post-layout simulation will ensure that the constraints created by pre-layout simulation are going to be met after introducing parasitics in the design.

On this section is presented POLO simulations for the different data paths lengths, classifying the paths in 3 different kinds: slow, medium and fast. Each of these paths are created by using different values for the "sel_delay" signal, choosing between a mixtures of buffers cells with high or smaller delay, plus a selection of high and low skew.

TABLE 5.2: SCHEMATIC VS LAYOUT DELAY COMPARISON FOR THE DIFFERENT DATA PATH CONFIGURATIONS

Delay Type	Stages	Data path Delay in Schematic	Data path Delay in Post-Layout
Slow	62	655ps	1.39ns
Medium	56	573ps	1.2ns
Fast	50	520ps	1.14ns

In Table 5.2, the different delays for the 3 paths configurations are presented,for both schematic and post-layout. It can be appreciated that POLO simulations show double the schematic delay. This is because now all the parasitic in the interconnections are being considered. Having these new delays in the circuit means the frequency needs to be scaled as well, otherwise the timing paths will present violations.

As it was explained in section 4.1.4, the voltage drop mitigation scheme has its own clock source; otherwise, the time-to-digital conversion will experience the clock-data compensation effect, causing bad results at the time to react to the voltage drop. In Table 5.3, the frequencies for the circuit and the mitigation scheme for both schematic and layout are compared.

TABLE 5.3: CLOCK CYCLE ADJUSTMENT BETWEEN SCHEMATIC AND POST-LAYOUT SIMULATIONS

| Circuit	Clock Cycle Schematic simulations	Clock Cycle POLO simulations
Tunable Data path	800 ps	1.55 ns
Voltage Drop Mitigation scheme	650 ps	1.3 ns

The tunable data path frequency was chosen in such a way the circuit have the same amount of slack in layout than in schematic. For example, in schematic the delay for the slow path is 655ps, and the clock cycle was 800ps, meaning there was about 150ps positive slack. Having 1.4ns delay

in the post-layout simulations means in order to get the same 150ps, clock cycle should be around 1.55ns.

For the mitigation scheme, the way the frequency was determined is by using a trial and error method, in which the cycle is large enough to have a similar code than schematic, but not too large to have aliasing through the delay line. It is worth to remember that the voltage threshold can be adjusted externally; this gives flexibility at the time to react to the voltage drop.

If the frequency is too fast, the VDM can start presenting aliasing in the MSBs, where the previous pulse has not traveled across all the registers are the code shows overlapping of 0s and 1s, very similar to the case where the voltage is too low to make an accurate sense. Figure 5.8 shows the case where the sampling frequency is too fast (910MHz) and aliasing start appearing in the VDM code.
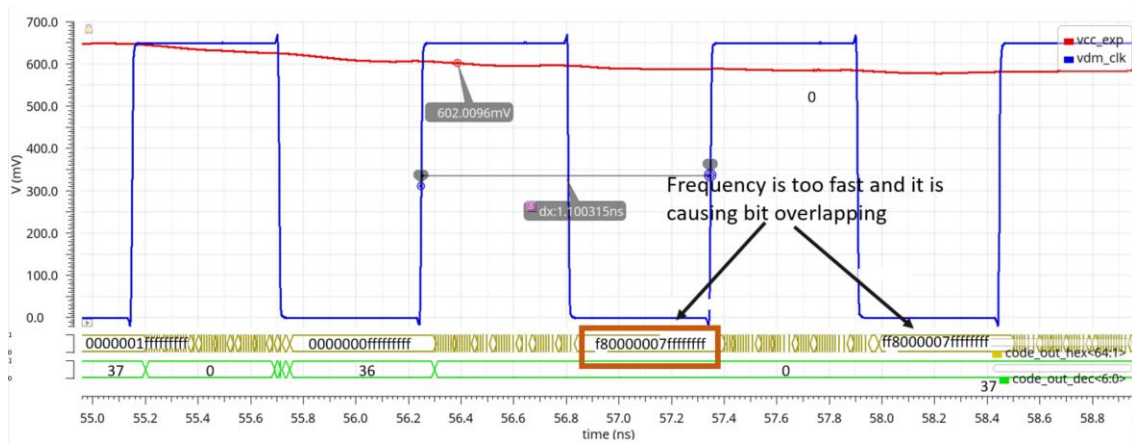
Figure 5.8: Aliasing in the VDM code due to fast sampling frequency @910MHz

Also, if the frequency is too slow, the voltage controller will not be able to react fast enough to counter the voltage drop. Figure 5. *9* shows how the VDM does not react fast enough and the circuit presents an error at its output; the system cannot tolerate have a voltage supply lesser than 600 mV, but since the sampling frequency is so slow, VDM will react when voltage is around 555mV, causing an error in the circuit.
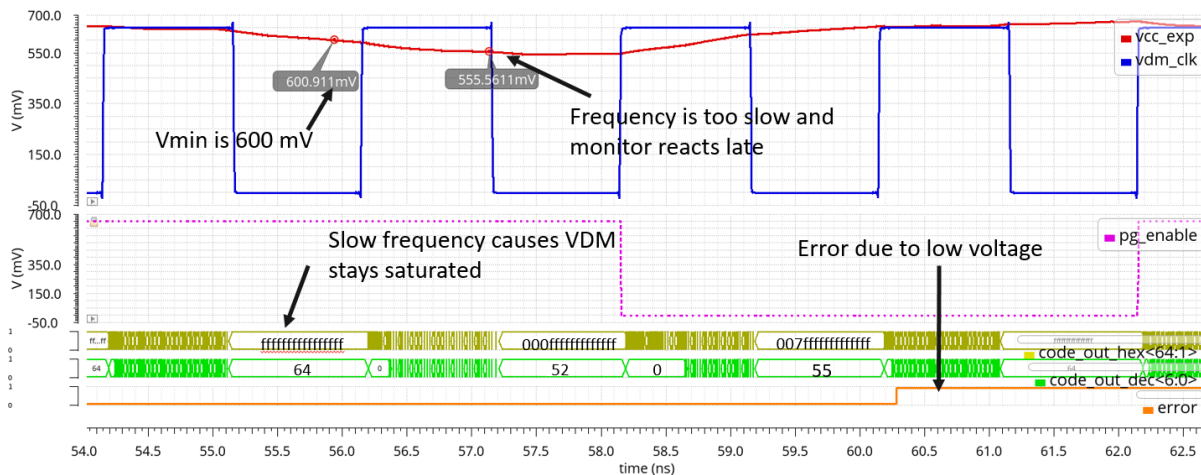
Figure 5. 9: VDM's late response at the time to have a slower frequency (500 MHz) than the load frequency (820 MHz)

5.3. Post-Layout Simulations

This section shows how the mitigation circuit is able to improve the voltage drop tolerance of the tunable data path by not failing under different current steps, including the parasitics extraction inherent to their physical implementation. Once again, these current steps mean the system presents a variation in the load by consuming more current from the voltage supply source. The baseline load current is 400μA, this is a little bit higher than the data path used in Chapter 3 since it was adapted to be tunable in the post-layout simulations, consuming more current at the time to have more cells. To show this improvement, we will present two types of tests: iso-frequency test and iso-current step test.

5.3.1. Iso-frequency Test

For this test, we need to understand the limitations of the circuit and figure what's the maximum current step the circuit can afford in a non-regulated configuration. Then, use the same current step in which the non-regulated circuit presents a failure, and check if the mitigation scheme can afford this amount of extra current. If it is able to afford this, the current step will be swept until the maximum point is reached. This procedure will used for the three kind of paths: slow, medium, and fast.

Slow Path without any Voltage Regulation

Figure 5.10 shows how the circuit in a non-regulated scheme can handle up to 40μA of extra current without causing any error. This corresponds to 10% of the whole current consumption. The voltage will drop up to 625mV, which is around 4% voltage drop. In Figure 5.11, it is observed how the current step of 50μA (12.5% of the load current) causes errors in the circuit when the voltage is recovering and the clock cycle is compressing due to the CDC effect. Here the voltage drops up to 620mV, which is around 5%. This means whatever change the load experience higher than 50μA, the circuit will fail.
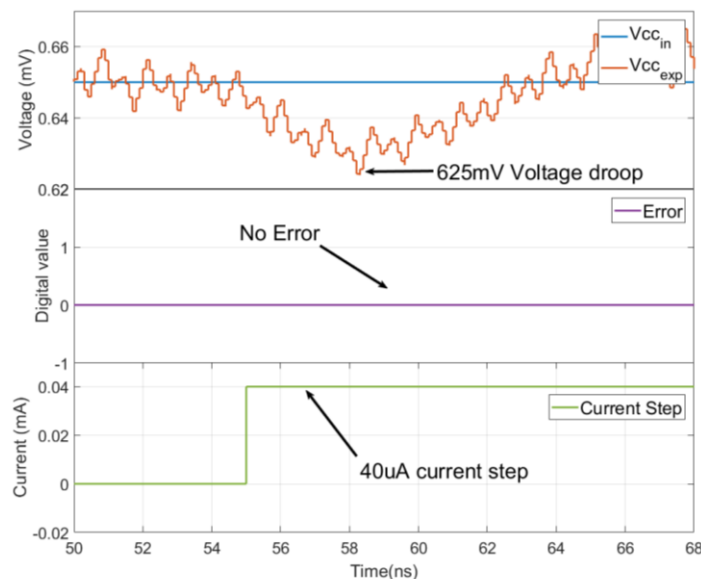
Figure 5.10: Maximum current step in slow path with no errors when no voltage regulation scheme is present
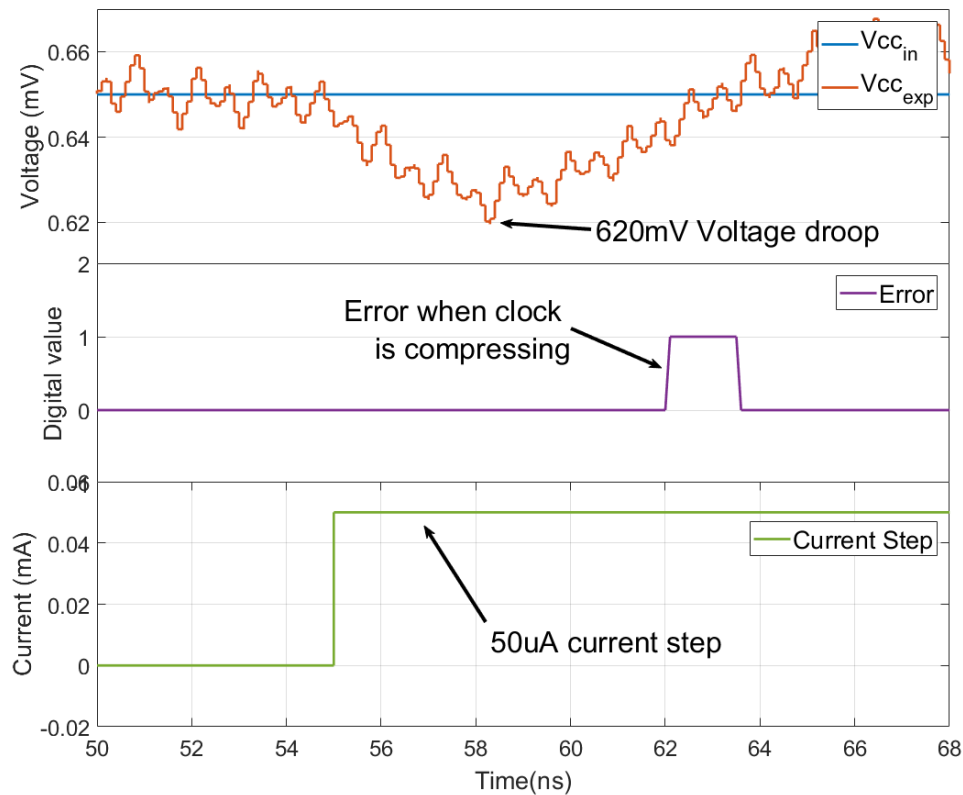
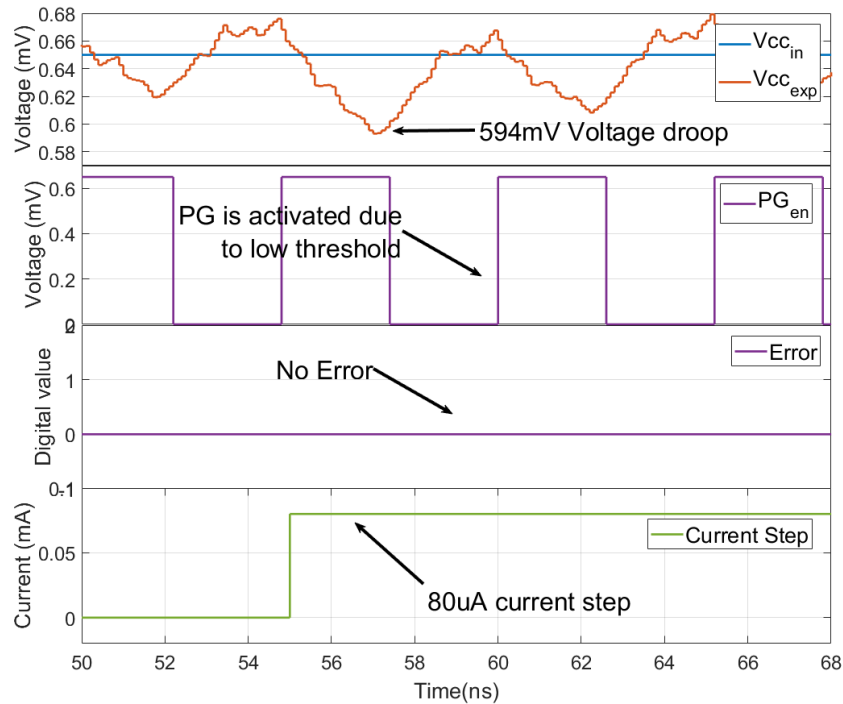**Figure 5.11:** Minimum current step causing errors in slow path when no voltage regulation scheme is present

Slow Path in a Voltage Regulated Configuration

Now the mitigation scheme is introduced to the system. The voltage threshold (PG code) is being set in such way the mitigation scheme will be able to control the voltage to avoid any failures. On this case, the voltage code will be 44.

Figure 5.12 shows how the circuit can afford up to 80µA of extra current without causing any error, this corresponds to 20% of the whole current consumption, being double the non-regulated test case. The voltage will drop up to 594 mV, which is around 9% voltage drop. Even this is lower than the non-regulated test case, the difference is the way the voltage is recovering that is steeper causing less clock period compression.

One thing to notice is the amount of positive slack for the slow path causes the voltage threshold is so small that variations in the supply make PG activated every time the voltage exceed the threshold.

Figure 5.12: Maximum current step in slow path with no errors when a voltage regulation scheme is present

Figure 5.13 shows how the current step of 90µA causes errors in the circuit when the voltage is over the Vcc_in nominal value[15]. Here the voltage drops up to 589mV, which is around 10%.
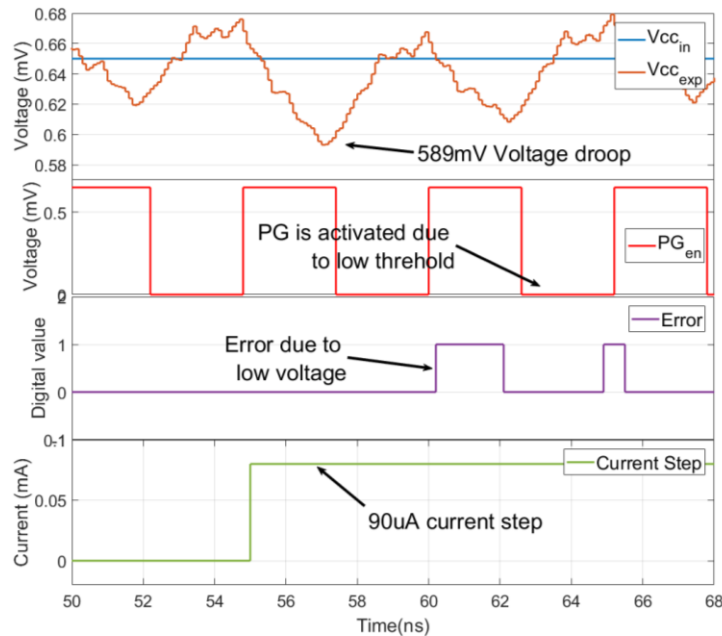
Figure 5.13: Minimum current step causing errors in slow path when a voltage regulation scheme is present

---

[15] This is also called overshoot

Medium Delay Path without any Voltage Regulation

First step is to determine the maximum current step the circuit can afford in a non-regulated configuration. Figure 5.14 shows how the circuit can handle up to 100µA of extra current without causing any error, this corresponds to 25% of the whole current consumption. The voltage will drops up to 600mV which is around 8% voltage drop. This is higher than the slow path because the timing path contains a higher positive timing slack, allowing the voltage to drop even further.
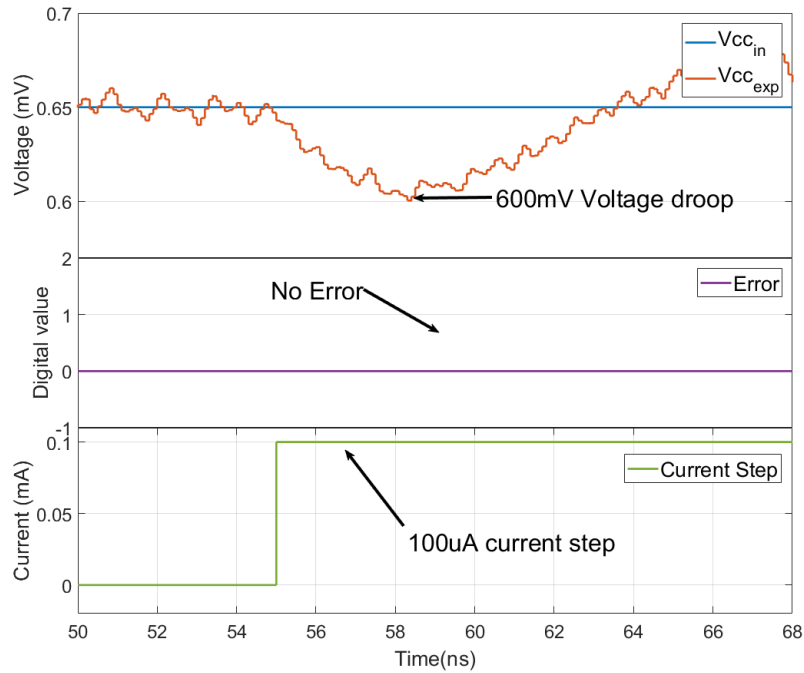
Figure 5.14: Maximum current step in medium delay path with no errors when no voltage regulation scheme is present

In Figure 5.15, it is observed how the current step of 120µA causes errors in the circuit when the voltage is recovering and the clock cycle is compressing. Here the voltage drops up to 593mV, which is around 9%.
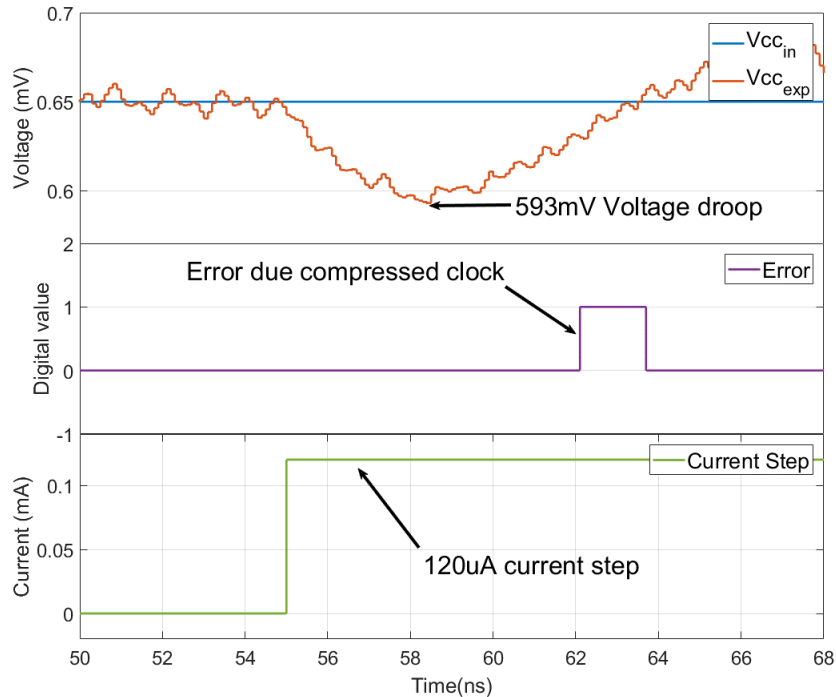
Figure 5.15: Minimum current step causing errors in medium delay path when no voltage regulation scheme is present

Medium Delay Path with Voltage Regulation

If the mitigation scheme is introduced to the system, we start testing how much extra current can the circuit handle before failing. The voltage threshold (PG code) is being set in such a way the mitigation scheme will be able to control the voltage to avoid any failures. On this case, the voltage code will be 41, which is around 610mV if we interpolate Table 4.2.

Figure 5.16 shows how the circuit can afford up to 165µA of extra current without causing any error; this corresponds to 41% of the whole current consumption, being 16% more than the non-regulated test case. The voltage will drops up to 586mV, which is around 10% voltage drop. Even this is lower than the non-regulated testcase, the difference is the way the voltage is recovering, which is steeper and this causes less clock period compression.

Figure 5.16: Maximum current step in medium delay path with no errors when a voltage regulation scheme is present

Figure 5.17, it is observed how the current step of 180μA causes errors in the circuit when the voltage is over the Vcc_in value. Here the voltage drops up to 581mV, which is around 11%.
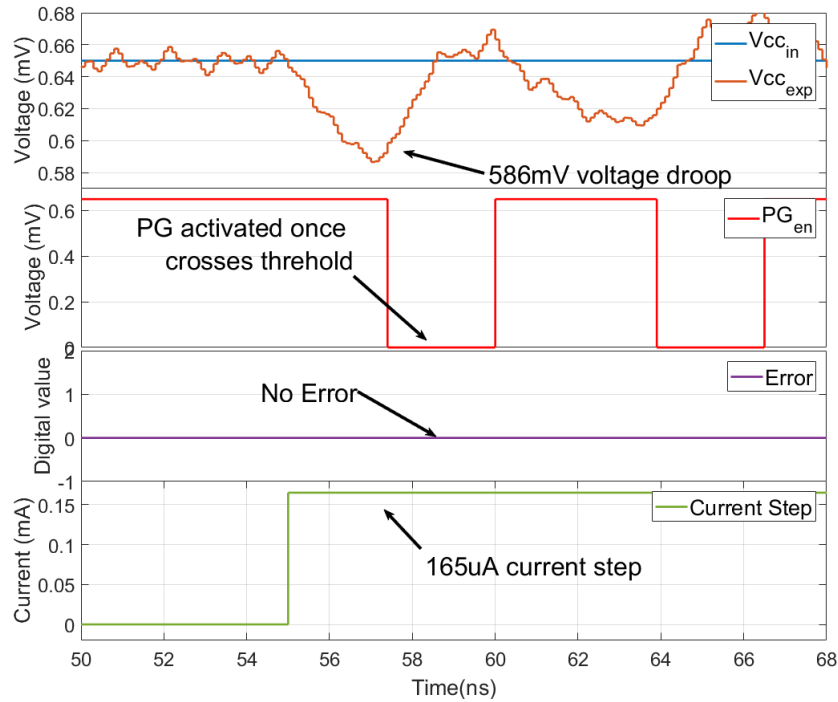

Figure 5.17: Minimum current step causing errors in medium delay path when no voltage regulation scheme is present
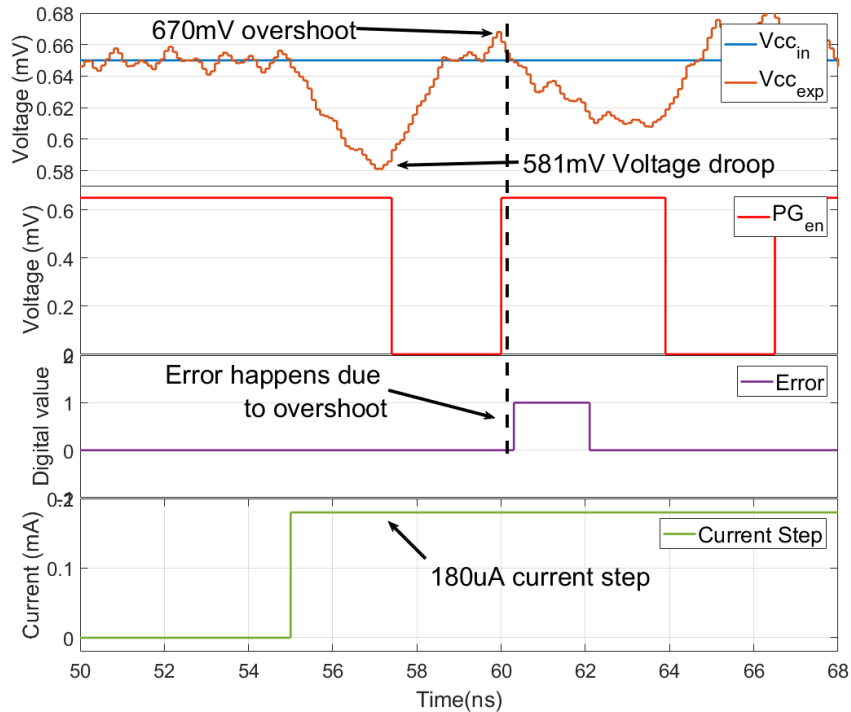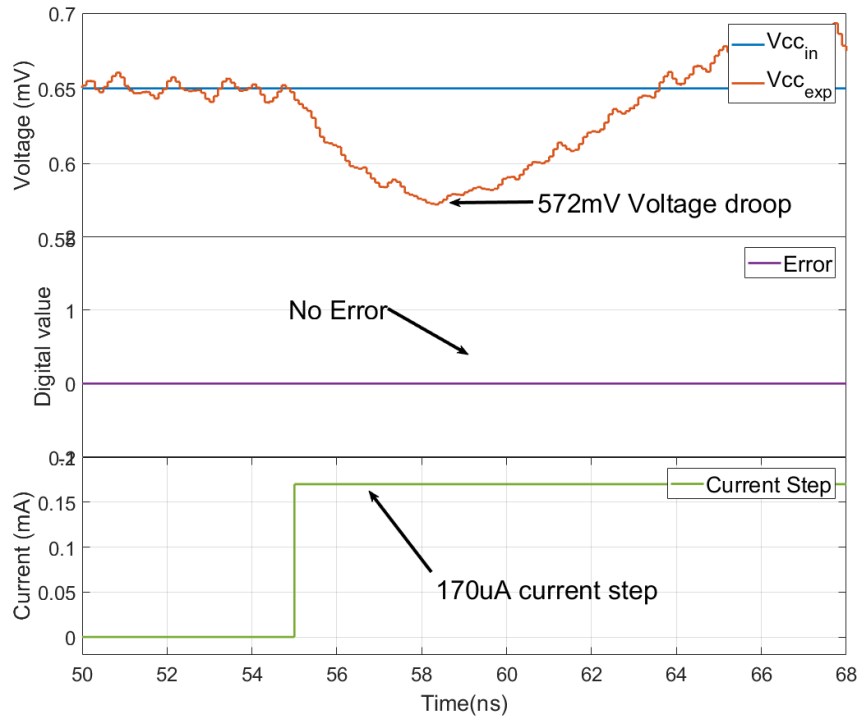
Fast Path without any Voltage Regulation

This scenario essentially will present the higher amount of extra current since it possess the higher positive timing slack, allowing the voltage to drop even more than the other two previous configurations. The maximum current step the circuit can afford in a non-regulated configuration it is showed in Figure 5.18, in which the circuit can handle up to 170µA of extra current without causing any error; this corresponds to 42.5% of the whole current consumption. The voltage will drop up to 572mV, which is around 12% voltage drop.

Figure 5.18: Maximum current step in fast path with no errors when no voltage regulation scheme is present

Figure 5.19 shows how the current step of 180µA causes errors in the circuit when the voltage is recovering and the clock cycle is compressing. Here the voltage drops up to 567mV, which is around 13%.

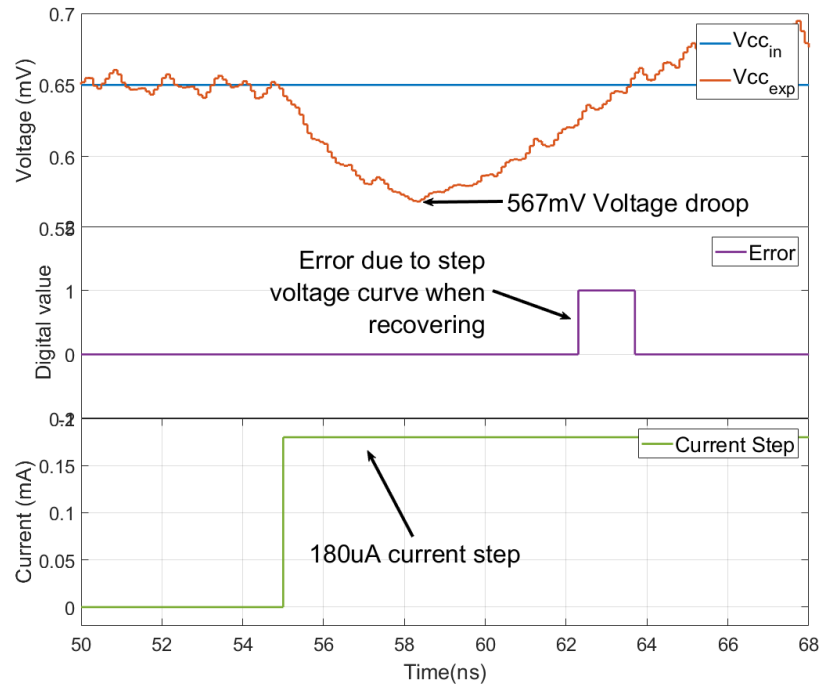**Figure 5.19:** Minimum current step causing errors in fast path when no voltage regulation scheme is present

Fast Path with Voltage Regulation

On this case, the voltage code will be 39, which is around 600mV. Probably the code can even be lower, however the issue is the voltage is dropping so fast that the jump between codes are too high, therefore we need to give some guardband to the PG_code. Lets remember the VDM is limited to the sampling frequency that was already set in Chapter 5.2.

Figure 5.20 shows how the circuit can afford up to 300µA of extra current without causing any error, this corresponds to 75% of the whole current consumption, being 43% more than the non-regulated testcase. The voltage will drop up to 538mV which is around 17% voltage drop. This is lower than the non-regulated testcase, the difference is the way the voltage is recovering that is steeper causing less clock period compression. The voltage takes around 4ns to get back to Vcc_in value (650mV).

79
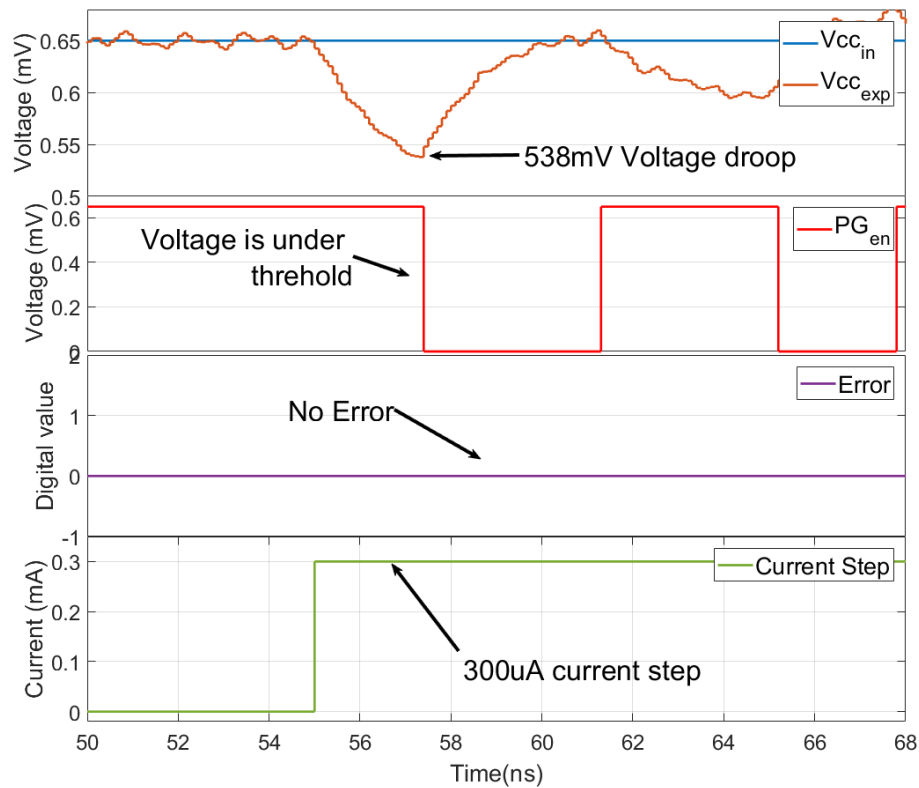
Figure 5.20: Maximum current step in fast path with no errors when a voltage regulation scheme is present.

Figure 5.21 presents a complete different scenario compared to the other to path delays when the mitigation scheme is being used. On this case, the problem is the controller does not work correctly, since the voltage is so small, causing aliasing in the VDM code, therefore, a misread in the code. Having this misread of the code, will cause "is_up" and "is_down" signals not behave correctly. Around 57 ns we can observed the overlap is happening, therefore, the controller thinks the voltage is going up when actually is going down, so even the threshold is being exceeded, the "PG_en" signal will not be activated. The minimum current step causing errors in fast path in a regulated configuration is 325µA, which is 81% of the actual load.

Figure 5.21: Minimum current step causing errors in fast path when no voltage regulation scheme is present
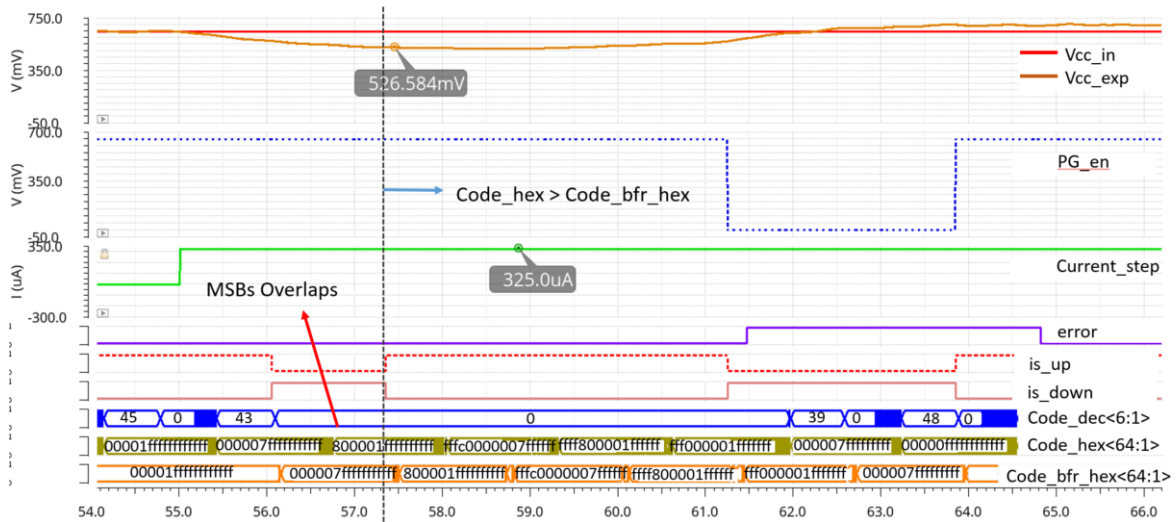
5.3.2. Iso-current Step Test

For the iso-current tests, the maximum current step in a regulated configuration without failures was used, applying this current step to the non-regulated circuit. Finally, we start sweeping the frequency to see the minimum frequency in which the non-regulated system won't fail. As in the previous sections, this is going to be followed for the three kind of paths: slow, medium and fast.

Slow Path
For the slow path, the clock cycle is being increased up to 1.65ns, meaning 100ps higher than the regulated configuration; this, in order to handle an 80 µA current step without any failures. With this new clock cycle, a voltage is allowed to drop up to 610mV.
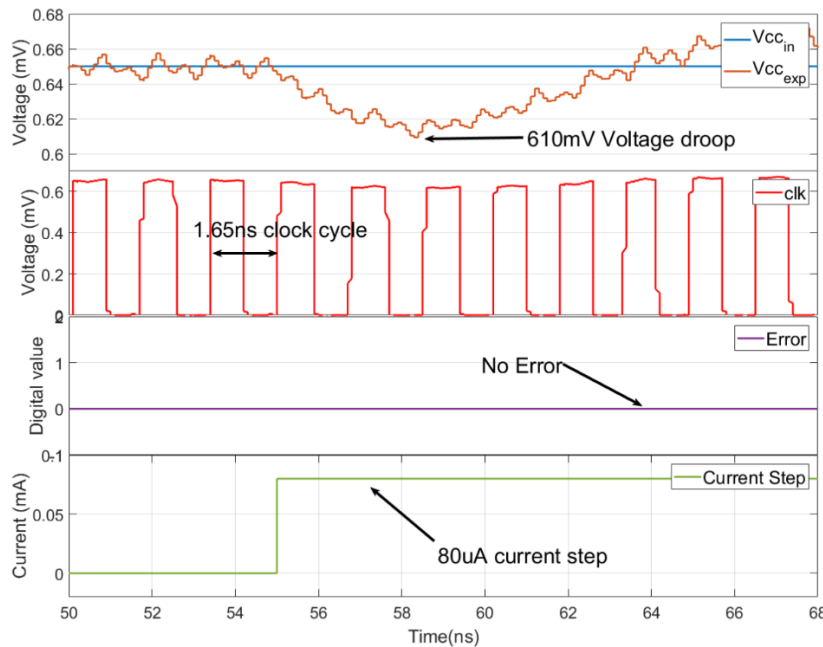
Figure 5.22: Minimum frequency for the slow path in a non-regulated configuration using regulated circuit max current step

Medium Delay Path

For the medium delay path, the clock cycle is being increased up to 1.65ns, being 100ps higher than the regulated configuration, this in order to afford a 165µA extra current step without errors. With this new clock cycle, voltage is allowed to drop up to 576mV.

Figure 5. 23: Minimum frequency for the medium delay path in a non-regulated configuration using regulated circuit max current step

Fast Path

For the fast path, the clock cycle is being increased up to 2ns, meaning 450ps slower than the regulated configuration, this in order to handle a 300 µA current step without any failures. With t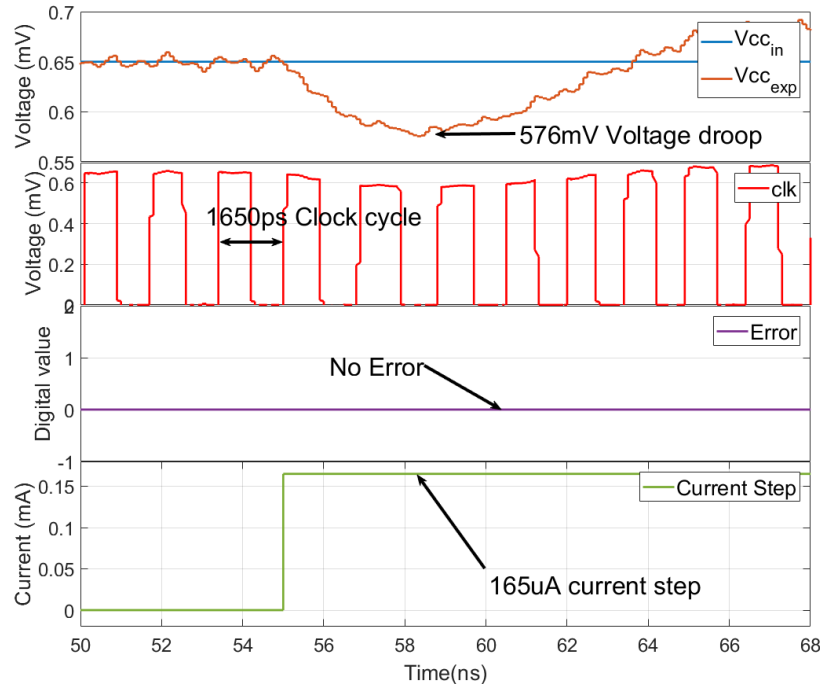his new clock cycle, voltage is allowed to drop up to 516mV. Again, this voltage is too low for the controller to work properly, but this is just a test of the non-regulated system.

Figure 5.24: Minimum frequency for the fast path in a non-regulated configuration using regulated circuit max current step
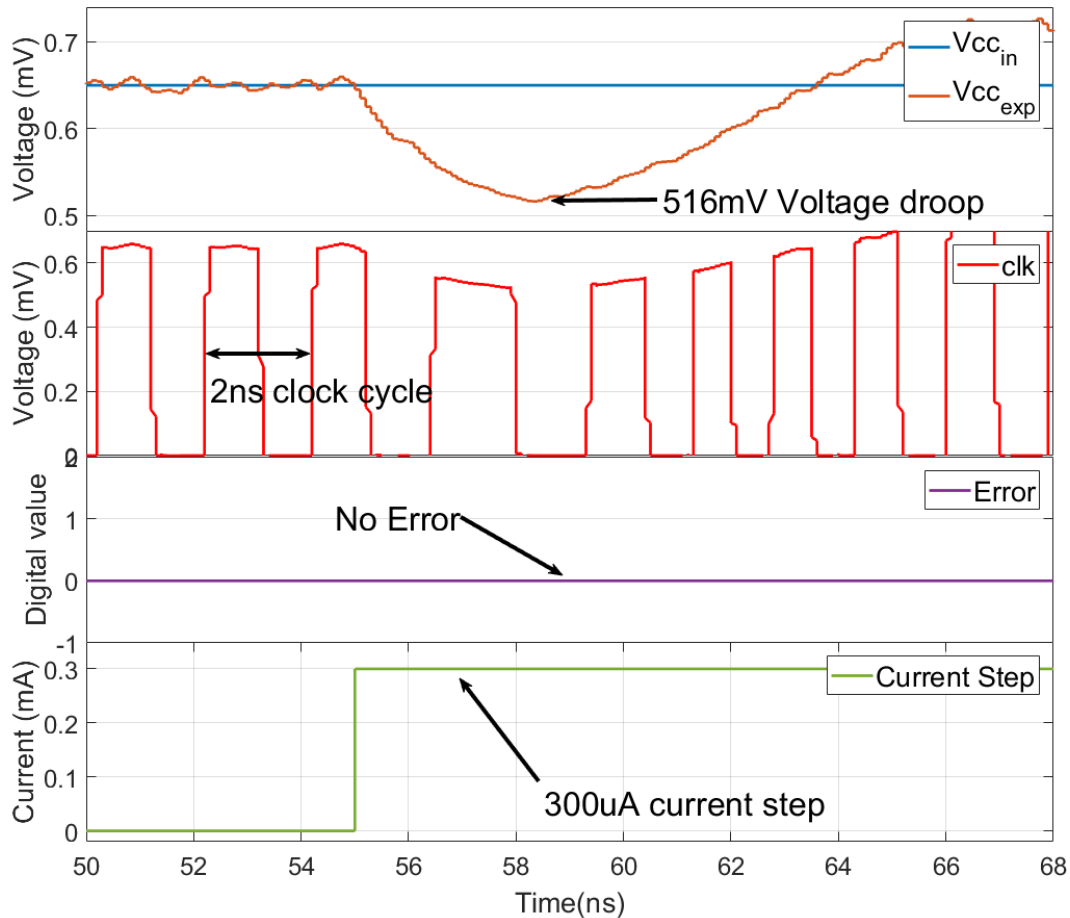
5.4. Summary of Results

This section presents a summary of the results of the above sections, highlighting the contributions of this work, which main objective is to show an improvement in the voltage drop guardband reduction at the time to design a circuit. Figure 5.25 presents an overview of the different guardband needed when a chip is being designed, like the intrinsic Vmin (which is technology process depended); also, aging of the circuit which is related to how fast the circuit responds at the BOL (beginning of life) versus EOF (end of Life), and finally the guardband that takes into account a voltage drop. A reduction in the voltage drop guardband is the ultimately goal for this work. This guardband reduction will exist when a voltage regulated system (Figure 5. 25(b)) presents less voltage drop guardband than non-regulated system (Figure 5. 25(a))
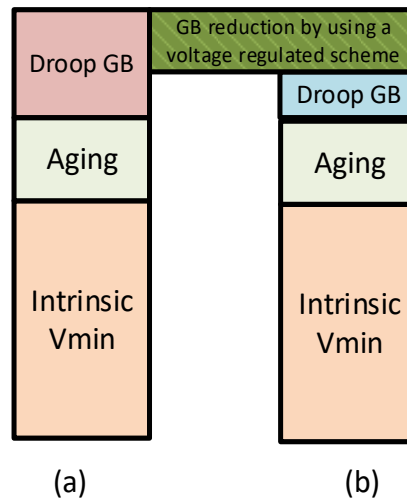
Figure 5. 25: Guardband distribution in a circuit design: (a) Voltage non-regulated circuit; (b) Voltage regulated circuit

In the below tables, the guardband reduction and improvement of the whole system is shown. In Table 5.4 the summary of the iso-frequency test performed above is presented. In Table 5.5, the summary corresponds tothe iso-frequency test performed.

TABLE 5.4: ISO-FREQUENCY TEST SUMMARY

Data path Type	Non-regulated Max current step	Regulated Max current step	Extra Current tolerance improvement
Slow	40µA (10%)	80µA (20%)	10%
Medium	100µA (25%)	165µA (41%)	16%
Fast	170µA (42.5%)	300µA (75%)	32.5%

TABLE 5.5: ISO-CURRENT STEP TEST SUMMARY

Data path Type	Regulated Max current step (645 MHz)	Max Freq. Non-Regulated system Iso-current step	Frequency improvement
Slow	80µA	605 MHz	6.2%
Medium	165µA	605 MHz	6.2%
Fast	300µA	500 MHz	22.4%

As it was shown, the proposed scheme increases the voltage drop tolerance of the system. Incorporating a scheme like this will bring some complexity to the power management portion of the design plus an overhead in terms of area of the circuit. In this case, the mitigation scheme area is 5 times bigger than the critical path. However, the data path area is only representing the critical path; in reality, the load will be much bigger than the proposed data path. Another consideration at the time to incorporate a system like this is the load frequency, since as it was seen, at least in the proposed VDM's configuration, frequency cannot be faster than 770 MHz.

Conclusions and Outlook

This work proposed a framework that enables Clock-Data-Compensation (CDC) analysis and mitigation, considering different voltage drop specs using a second order power grid model. CDC effect have been analyzed through different test scenarios. These test scenarios take into account different clock tree lengths and also a non-clock tree scenario. A higher CDC was observed when the clock tree sensitivity is closer to the data path sensitivity, and a very short clock path makes the clock period modulation effect weaker. Also, a system that enables adaptive voltage scaling to mitigate CDC was demonstrated at post-layout simulation level, increasing performance and energy efficiency in the circuit by enhancing the voltage drop tolerance. An in-situ voltage drop monitor (VDM) was developed to detect and react to the voltage drop in a very fast way. The VDM creates a code depending on the voltage by doing a time-to-digital conversion. The circuit allows having a multi-threshold operation plus voltage drop direction detection, in which the controller can react depending on the voltage level. In order to have a better performance of the VDM, the clock source needs to be isolated from the system clock, otherwise the VDM will be affected by the CDC effect, causing an inaccurate code when a voltage drop is occurring.

Through a voltage controller that connects different power gates, charge is being injected to the circuit every time the voltage supply exceed this threshold. The voltage controller turns on the power gates depending on the voltage supply value and if it is exceeding the low or high limits of the controller. The power gates need to be connected to a higher or same level voltage supply than the circuit in order to regulate the voltage level.

Finally, a tunable data path was implemented in order to have different timing paths when performing tests. A layout implementation of the circuits was done in order to perform post-layout simulations and create an accurate model closer to reality. The post-layout simulations show a 10% (slow path), 16% (medium delay path) and 32.5% (fast path) extra current tolerance improvement in the circuit using the voltage drop mitigation scheme compared a circuit without it. Additionally, it show 6.2% (slow path), 6.2% (medium delay path) and 22.4% (fast path) frequency improvement in the circuit using the voltage drop mitigation scheme in comparison to the reference case without it. The proposed VDM comes with extra area overhead of $360\,\mu m^2$. Its response is limited by its maximum frequency, which is 770 MHz in this study. This limits the voltage controller reaction if the load frequency is much higher than this, and also its response to low voltages, since it can be a misread of the voltage code, causing incorrect behaviors of the voltage controller.

On this work the voltage mitigation scheme was designed using an adaptive voltage scaling configuration, however, this scheme opens the possibility of using other approaches like adaptive frequency scaling. In future works, it will be interesting to do a comparison of the advantages and disadvantages for both AVS and AFS schemes. Since CDC effect depends on the length of the clock, other future work can be related to investigate the circuit when the clock is managed in a different voltage in order to reduce power consumption at the time add extra stages in the clock tree. Other topic of interest for future research might be to perform a different voltage control of the power gates, in which depending of the VDM code value, different number of power gates can be turned on or off.

References

[1] Bowman, K. A., Tokunaga, C., Karnik, T., De, V. K., & Tschanz, J. W. (2012). A 22nm dynamically adaptive clock distribution for voltage drop tolerance. VLSI Circuits (VLSIC), 2012 Symposium on, 48(4), 94–95.

[2] S. Das et al., "A self-tuning DVS processor using delay-error detection and correction,"IEEE J. Solid-State Circuits, pp. 792–804, Apr. 2006.

[3] S. Das et al., "Razor II: In situ error detection and correction for PVT and SER tolerance,"IEEE J. Solid-State Circuits, pp. 32–48, Jan. 2009.

[4] K. A. Bowman et al., "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," IEEE J. Solid-State Circuits, pp. 49–63, Jan. 2009.

[5] K. A. Bowman et al., "A 45 nm resilient microprocessor core for dynamic variation tolerance,"IEEE J. Solid-State Circuits, pp. 194–208, Jan. 2011.

[6] J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits,"IEEE Trans. Device Mater. Rel., vol. 1, pp. 43–59, Mar. 2001.

[7] T. Ichikawa and M. Sasaki,"A new analytical model of SRAM cell stability in low-voltage operation," IEEE Trans. Electron Devices, vol. 43, pp. 54–61, Jan. 1996.

[8] Saint-Laurent, M., & Swaminathan, M. (2004). "Impact of Power-Supply Noise on Timing in High-Frequency Microprocessors". IEEE Transactions on Advanced Packaging, 27(1), 135–144.

[9] Wong, K. L., Rahal-arabi, T., Ma, M., & Taylor, G. (2006). "Enhancing Microprocessor Immunity to Power Supply Noise With Clock-Data Compensation". IEEE Journal of Solid-State Circuits, 41(4), 749–758.

[10] Jiao, D., Gu, J., & Kim, C. H. (2010). "Circuit design and modeling techniques for enhancing the clock-data compensation effect under resonant supply noise". IEEE Journal of Solid-State Circuits, 45(10), 2130–2141.

[11] T. Rahal-Arabi, G. Taylor, M. Ma, and C. Webb, "Design and validation of the Pentium III and Pentium 4 processors power delivery," inSymp. VLSI Circuits Dig., Jun. 2002, pp. 220–223.

[12] S. Pant and E. Chiprout, "Power grid physics and implications for CAD," inProc. Design Automation Conf. (DAC), Jul. 2006, pp. 199–204.

[13] E. Hailu, D. Boerstler, K. Miki, J. Qi, M. Wang, and M. Riley, "A circuit for reducing large transient current effects on processor power grids," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2006, pp. 2238–2245.

[14] N. Kurd, P. Mosalikanti, M. Neidengard, J. Douglas, and R. Kumar, "Next generation Intel core micro-architecture (Nehalem) clocking," IEEE J. Solid-State Circuits, vol. 44, no. 4, pp. 1121–1129, Apr. 2009.

[15] J. M. Rabaey, A. Chandrakasan and B. Nikolic, Digital Integrated Circuits A Design Perspective, 2003.

[16] N. Kurd, P. Mosalikanti, M. Neidengard, J. Douglas, and R. Kumar, "Next generation Intel core micro-architecture (Nehalem) clocking," IEEE J. Solid-State Circuits, vol. 44, no. 4, pp. 1121–1129, Apr. 2009.

[17] T. Rahal-Arabi, G. Taylor, J. Barkatullah, K. L. Wong, and M. Ma, "Enhancing microprocessor immunity to power supply noise with clock/data compensation," inSymp. VLSI Circuits Dig., Jun. 2005, pp. 16–19.

[18] K. L. Wong, T. Rahal-Arabi, M. Ma, and G. Taylor, "Enhancing microprocessor immunity to power supply noise with clock-data compensation,"IEEE J. Solid-State Circuits, vol. 41, no. 4, pp. 749–758, Apr. 2006.

[19] D. Jiao, J. Gu, P. Jain, and C. Kim, "Enhancing beneficial jitter using phase-shifted clock distribution," inProc. IEEE Int. Symp. Low Power Electronics and Design (ISLPED), Aug. 2008, pp. 21–26.

[20] N. Kurd, J. Barkatullah, and P. Madland, "Adaptive frequency clock generation system," US Patent 7,042,259 B2, May 9, 2006.

[21] N. A. Kurd, J. S. Barkarullah, R. O. Dizon, T. D. Fletcher, and P.D. Madland, "A multigigahertz clocking scheme for the Pentium 4 microprocessor,"IEEE J. Solid-State Circuits, vol. 36, no. 11, pp. 1647–1653, Nov. 2001

[22] Kim, S. T., Shih, Y., Mazumdar, K., Jain, R., Ryan, J. F., Tokunaga, C… De, V. "Enabling Wide Autonomous DVFS in a 22 nm Graphics Execution Core Using a Digitally Controlled Fully Integrated Voltage Regulator", 1–13, March 2015.

[23] K. K. Rangan, G. -Y. Wei, and D. Brooks, "Thread motion: Fine-grained power management for multi-core systems," in Proc. 36th Annu. Int. Symp. Computer Architecture (ISCA), Jun. 2009.

[24] Y. Okuma et al., "0.5-V input digital LDO with 98.7% current efficiency and 2.7-mA quiescent current in 65 nm CMOS," inProc. IEEE CICC, 2010, pp. 313–326.

[25] K. Hirairiet al., "13% power reduction in 16b integer unit in 40 nm CMOS by adaptive power supply voltage control with parity-based error prediction and detection (PEPD) and fully integrated digital LDO," in IEEE ISSCC Dig. Tech. Papers, 2012, pp. 486–487.

[26] S. Gangopadhyay et al., "A 32 nm embedded, fully-digital,phase-locked low dropout regulator for fine grained power management in digital circuits," IEEE J. Solid-State Circuits, vol. 49, no. 11, pp. 2684–2693, Nov. 2014.

[27] S. B. Nasiret al., "A0.13 m fully digital low-dropout regulator with adaptive control and reduced dynamic stability for ultra-wide dynamic range," in IEEE ISSCC Dig. Tech. Papers, 2015, pp. 98–99.

[28] Z. Toprak-Denizet al., "Distributed system of digitally controlled microregulators enabling per-core DVFs for the Power8 microprocessor," in IEEE ISSCC Dig. Tech. Papers, 2014, pp. 98–99.

[29] R. J. Millikenet al., "Full on-chip CMOS low-dropout voltage regulator," IEEETrans.CircuitsSyst.I,Reg.Papers, vol. 54, no. 9, pp. 1879–1890, Sep. 2007.

[30] Kevin T. Tang and Eby G. Friedman, Fellow, IEEE, "Simultaneous Switching Noise in On-Chip CMOS Power Distribution Networks" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 10, No. 4, August 2002.

[31] Patrik Larsson, "di/dt Noise in CMOS Integrated Circuits". Analog Integrated Circuits and Signal Processing, 14, 113–129. 1997

[32] Istvan Novak, Fellow, IEEE, "Lossy Power Distribution Networks With Thin Dielectric Layers and/or Thin Conductive Layers". IEEE Transactions On Advanced Packaging, Vol. 23, No. 3, August 2000.

[33] Patrik Larsson. Larsson, "Power supply noise in future IC's: a crystal ball reading". Custom Integrated Circuits, 1999. Proceedings of the IEEE 1999.

[34] [H. B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI, Addison-Wesley, 1990

[35] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," IEEE J. Solid-State Circuits, vol. 25, pp. 584–594, Apr. 1990.

[36] N. H. E. West and K. Eshraghian, Principles of CMOS VLSI Design: a System Perspective, 2nd ed. Reading, MA: Addison-Wesley, 1993, ch. 8.

[37] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," Proc. IEEE, vol. 89, pp. 490–504, Apr. 2001.

[38] Microsemi, "Simultaneous Switching Noise and Signal Integrity", 2012

[39] A. Wang, S. Naffziger, "Adaptive Techniques for Dynamic Processor Optimization. Springer Science+Business Media, LCC 2008

[40] L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc, and T. Roy, "Power distribution system design methodology and capacitor selection for modern CMOS technology," IEEE Trans. Adv. Packag., vol. 22, no. 3, pp. 284–291, Aug. 1999.

[41] M. Budnik and K. Roy, "A power delivery and decoupling network minimizing ohmic loss and supply voltage variation in silicon nanoscale technologies," IEEE Trans. VLSI, vol. 14, no. 12, Dec. 2006, pp. 1336–1346.

[42] Weste, N. H. E. and Harris, D. 2008. CMOS VLSI Design: A Circuits and Systems Perspective, 4th ED. Addison-Wesley.

[43] V. Gutnik and A. Chandrakasan, "Embedded power supply for low-power DSP" IEEE Trans. VLSI Syst. vol. 5, no 4, pp 425-435, 1997

[44] J. Tschanz et al., "Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging", IEEE ISSCC Dig. Tech. Papers, Feb. 2007.

[45] Seevinck, E, List, F, Lohstroh, J, "Static noise margin analysis of MOS SRAM cells," IEEE Journal of Solid-State Circuits, Vol. 22, No. 5, pp. 748–754, October 1987.

[46] Clark, L, et al., "An embedded microprocessor core for high performance and low power applications," IEEE Journal of Solid-State Circuits, Vol. 36, No. 11, pp. 498–506, November 2001.

[47] S. Samaan, "The Impact of Device Parameter Variations on the Frequency and Performance of VLSI Chips," ICCAD, 7–11 Nov 2004, pp. 343–346.

[48] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, and V. Pokala, "A Distributed Critical-Path Monitor for a 65nm High-Performance Microprocessor," ISSCC, 11–15 Feb 2007, pp. 398–399.

[49] M. Elgebaly and M. Sachdev, "Variation-Aware Adaptive Voltage Scaling System," IEEE Transactions on VLSI Systems, vol. 15, no. 5, May 2007, pp. 560—571.

[50] G. Wolrich, E. McLellan, L. Harada, J. Montanaro, and R. Yodlowski, "A high performance floating point coprocessor," IEEE Journal of Solid-State Circuits, Volume 19, Issue 5, October 1984.

[51] S. Lu, "Speeding up processing with approximation circuits," IEEE Micro Top Picks, pp. 67–73, 2004

[52] D. Ernst et al., "Razor: a low-power pipeline based on circuit-level timing speculation," *Proc. Intl.Symp. Microarchitecture*, Dec. 2003, pp. 7–18.

[53] S. Das et al., "A self-tuning DVS processor using delay-error detection and correction," JSSC, vol. 41, no. 4, Apr. 2006, pp. 792–804.

[54] K. Shu ; E. Sanchez ; J. Silva; S.H.K. Embabi "A 2.4-GHz monolithic fractional-N frequency synthesizer with robust phase-switching prescaler and loop capacitance multiplier" IEEE Journal of Solid-State Circuits vol. 38, no. 6, pp. 866 - 874, Jun. 2003

[55] Hugh Mair, Ericbill Wang, Alice Wang "A 10nm FinFET 2.8GHz Tri-Gear Deca-Core CPU Complex with Optimized Power-Delivery Network for Mobile SoC Performance" in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2017, pp. 56–58.

Appendix A

VDM considerations: Delay Line problem

Initially delay line was inverted based (Figure A.1), fall and rise times differ slightly, but the sum of all this uncertainty will cause a difference in the MSB bits where the code is actually read by the controller. This issue is only seen in POLO simulations and not Schematic simulations:
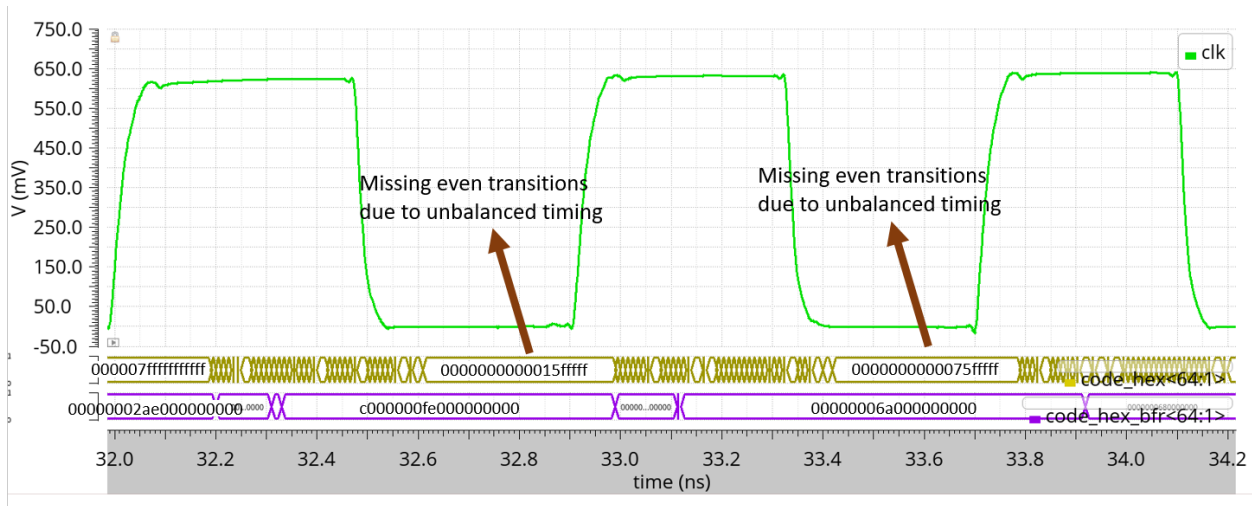
Inverter based delay line:
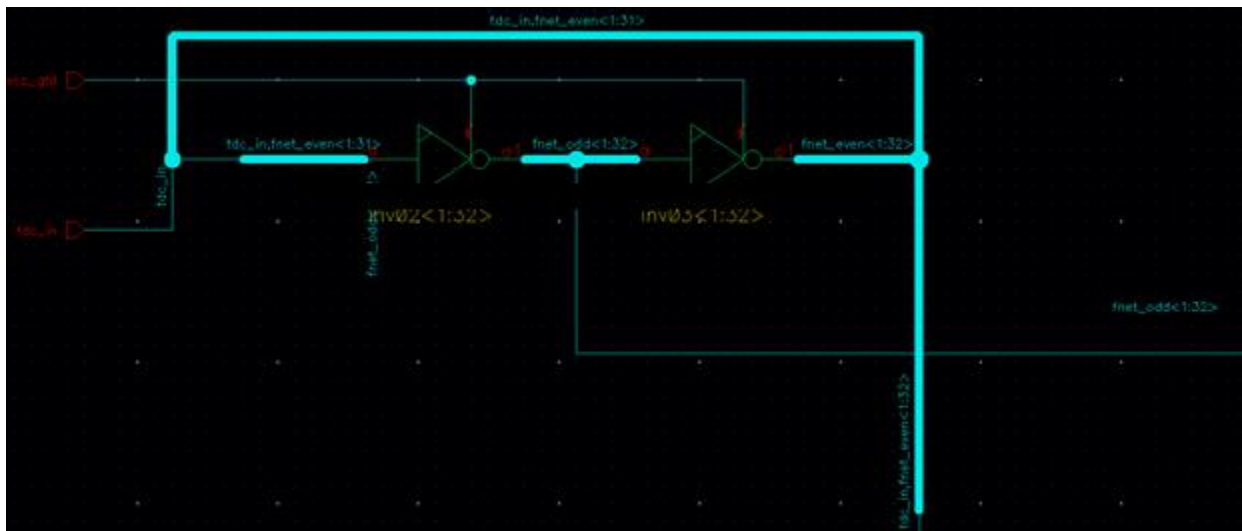
Figure A.1: Inverter based delay line

Not straight pattern of 0's and 1's. E.g.: 00000001**010**11111111

The issue is appreciated in Figure A.2, where the MSB bits are intercalated

Figure A.2: Missing even transitions in VDM code using inverters

To solve this, it was changed to a Buffer based Delay line. This comes with the penalty of the VDM will take more time to calculate the error, since the signal will take more time to travel from the LSB to the MSB. But, using buffer we are ensuring both fall and rise are the same value, cause a straight pattern of 0's and 1's. Figure A.3 shows this:
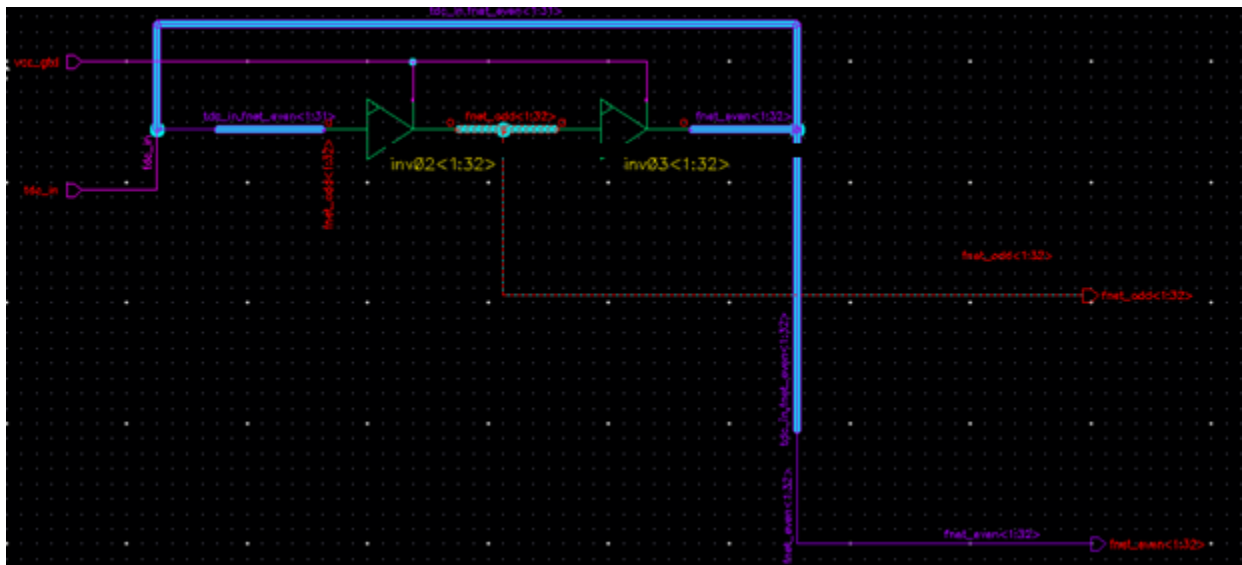

Figure A.3: Buffer based delay line

Appendix B

VDM Considerations: Hold Timing Issues for "Previous Code" Calculation

At the time to generate the code, and "previous code" for the Controller to determine if the voltage is going up or down, the problem is there were hold violations between the latches of the "present code" and flops of the "previous code"
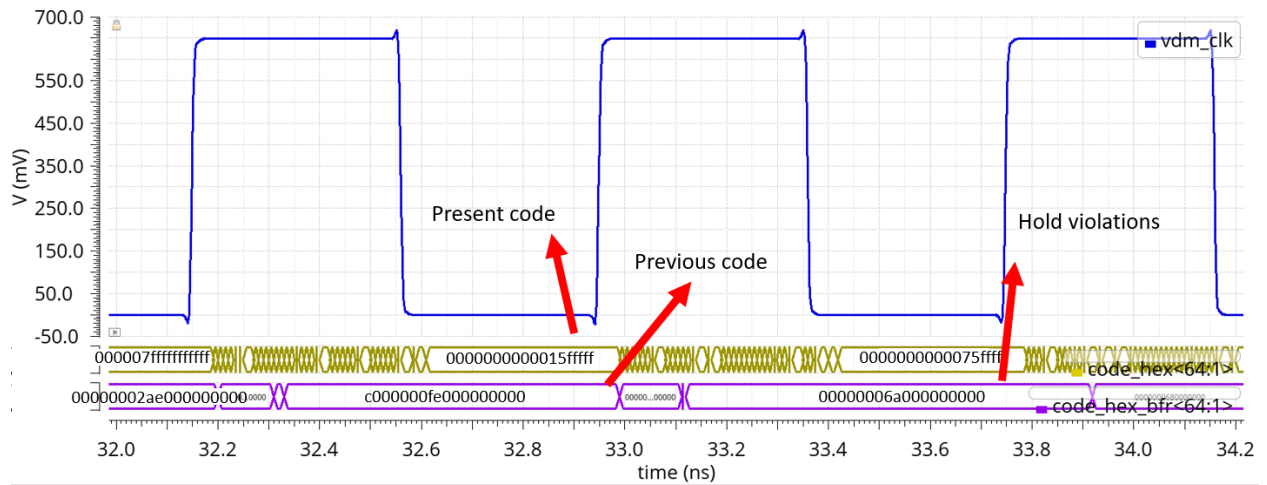
In the figure you can see there is no buffer going to the input of the FF as is presented in Figure B.1
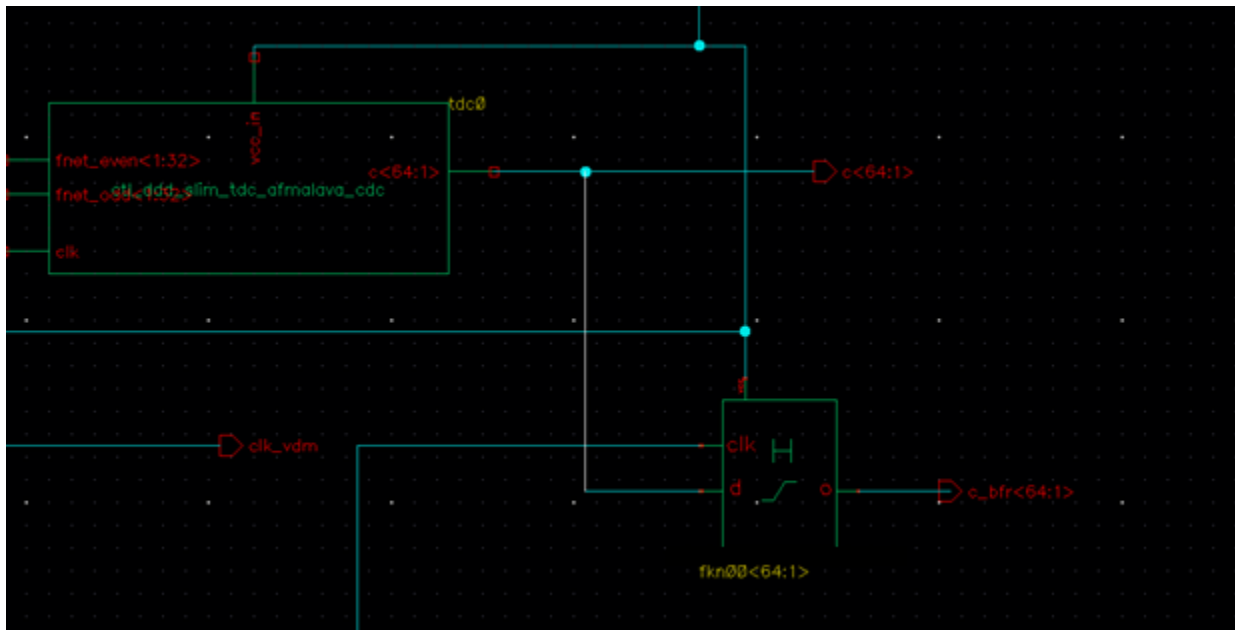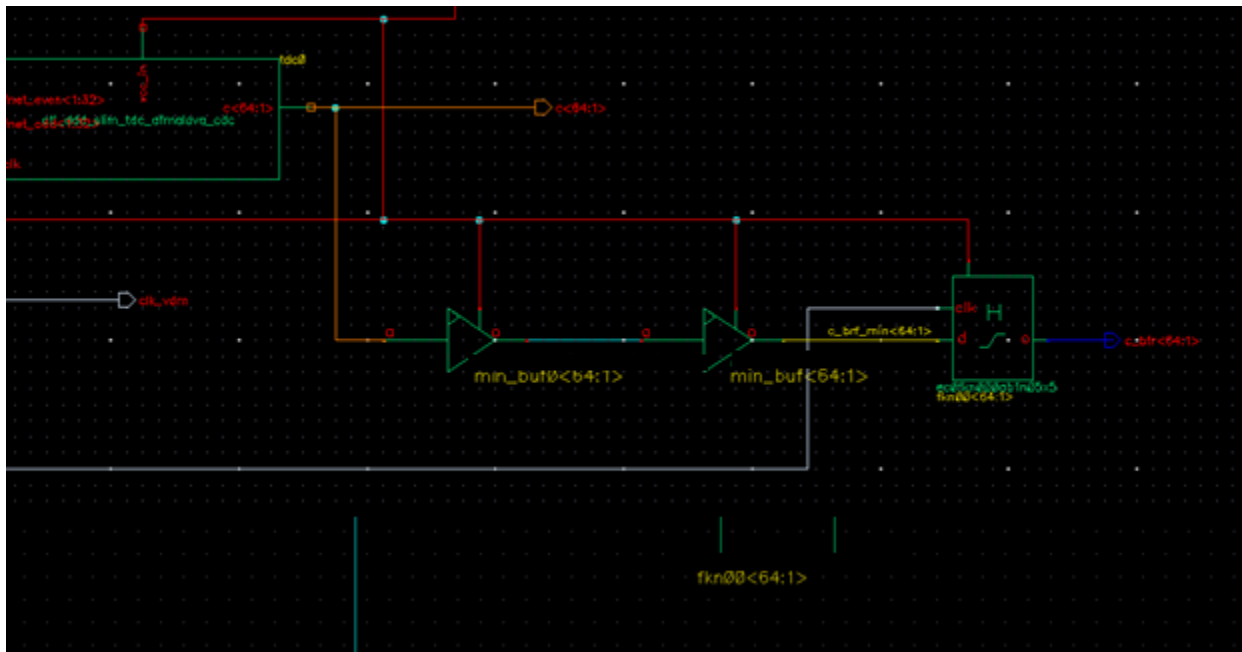

Figure B.1: VDM without hold buffers

This cause hold violations for the "previous code" generation as is shown in Figure B.2

Figure B.2: Hold violations in previous code for the VDM

If we insert enough buffer, the issue will be solved as is shown in Figure B.3:

Figure B.3: VDM with hold buffers

Appendix C

VDM Behavior under Different Voltage Drop Frequencies:

So far, the system has deal with voltage drop of 100MHz. According to Figure 3.3, the goal is to create a mitigation scheme which responds to high frequency voltage drops (which are the first drop during a voltage drop event). These drops goes from 50MHz to 250MHz, therefore, it's important to know the behavior of the scheme under different frequencies.

In Figure C.1 and Figure C.2, can be found voltage drops with frequencies of 50MHz and 250MHz respectively. For the 50MHz, the code is changing since the voltage level from clock edge to the next edge is different, however, this change is not big compare to the 100MHz drop. This because the frequency of the drop is slow respect to the current clock frequency (1.5 GHz). Now, for a faster frequency there is a change in the code as well, but only for a few cycles since drop happens really fast respect to the clock period.
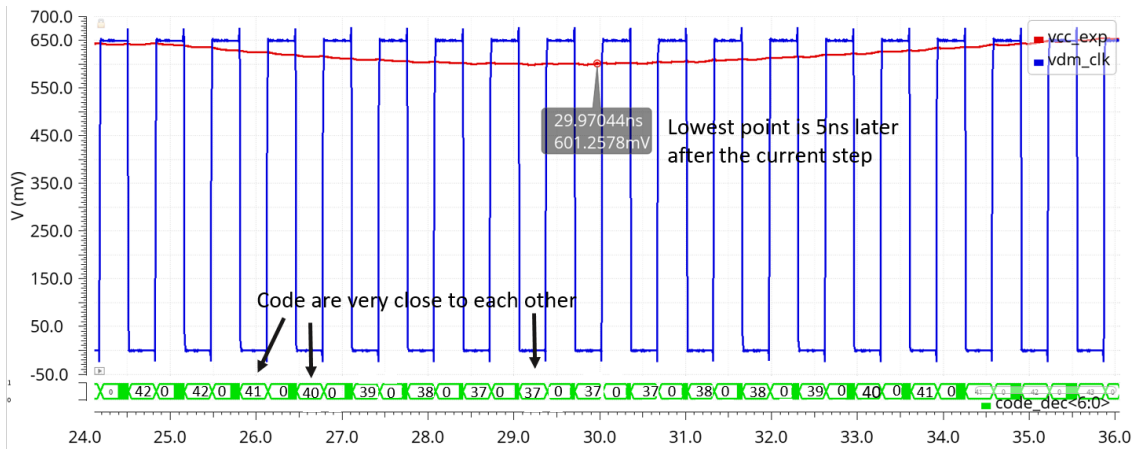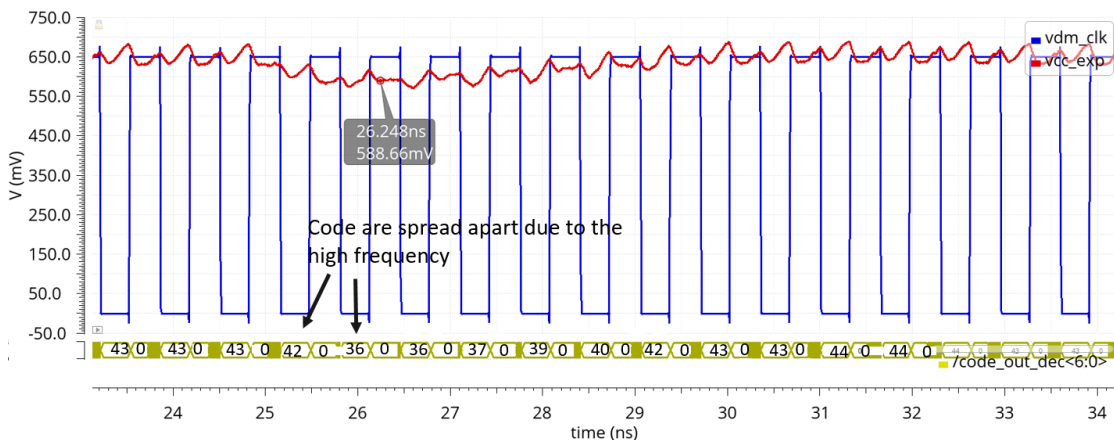

Figure C.1: Voltage drop of 50 MHz frequency

Figure C.2: Voltage drop of 250 MHz frequency

94

Appendix D

Clock Swing Issues for VDM's Clock Tree

At the time to perform Post-layout simulations on the voltage drop controller while the rest of the blocks are in the schematic view.

The first issue was the clock signal, which doesn't gives a full swing across the simulation. In Figure D.1, the yellow wave is the schematic (ideal) signal and the red one is the extracted (layout) signal.
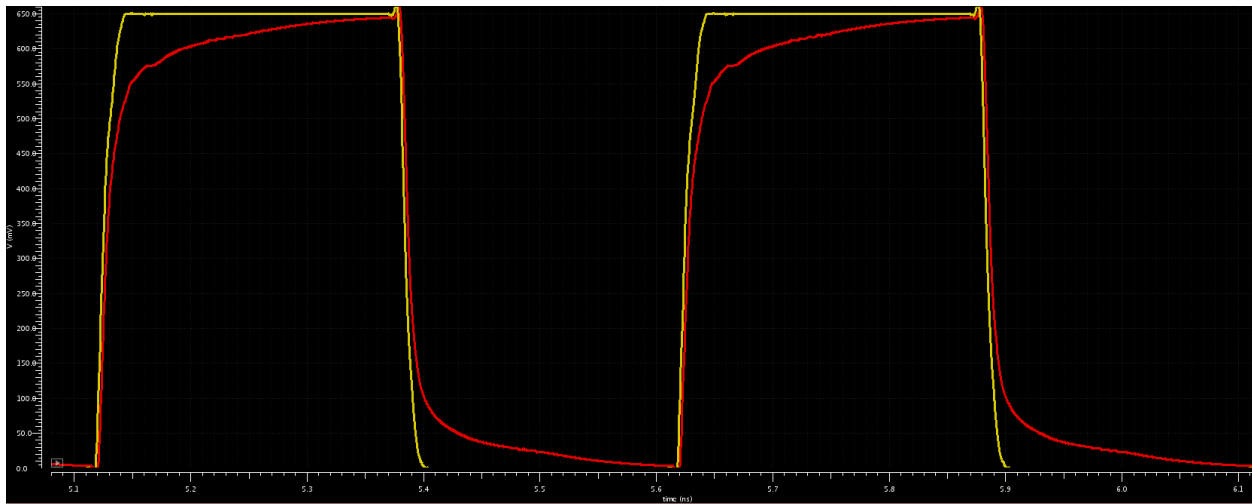

Figure D.1: VDM's clock full swing problem (POLO simulations)

Digging into this problem, we realized the clock driver strength was not enough to hold all the sinks at the final node of the clock tree. The clock tree looks as in Figure D.2:

(a)

(b)
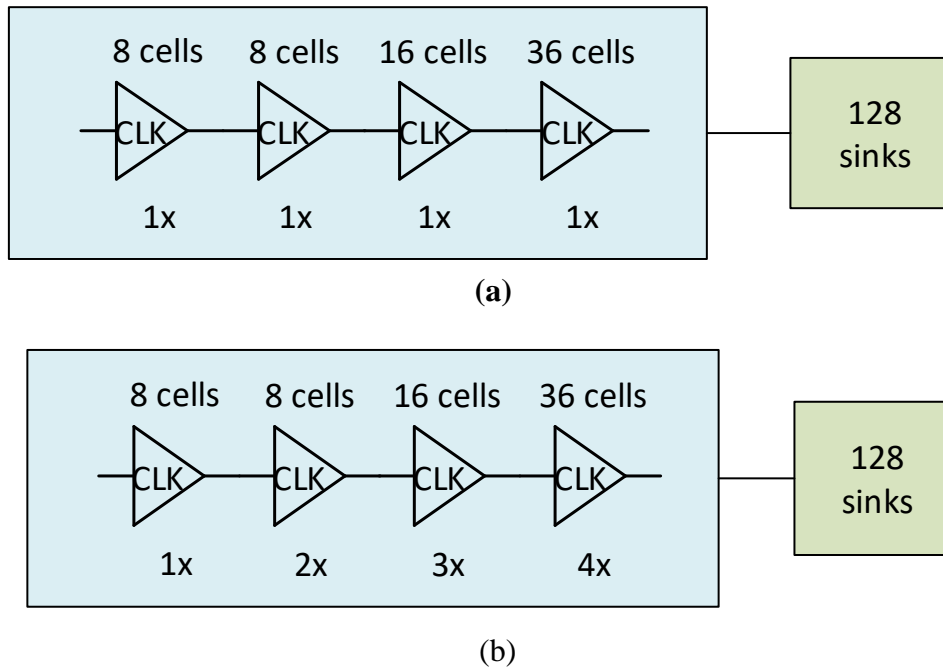
Figure D.2: VDM's Clock tree: (a) Initial approach; (b) New approach to give full swing

In the Figure D.2, we have the previous clock tree in which all the clock cells are the same size. Having 1x cells to drive 128 sinks, is too low small for such capacitance, that's why we have the previous waveform. Now, if we incrementally upsize the different stages in the clock tree, we are going to improve the drive strength of the final stage of the clock tree. Using this new structure will have the results in Figure D.3, in which the schematic wave is the yellow and the post-layout wave is the green one.
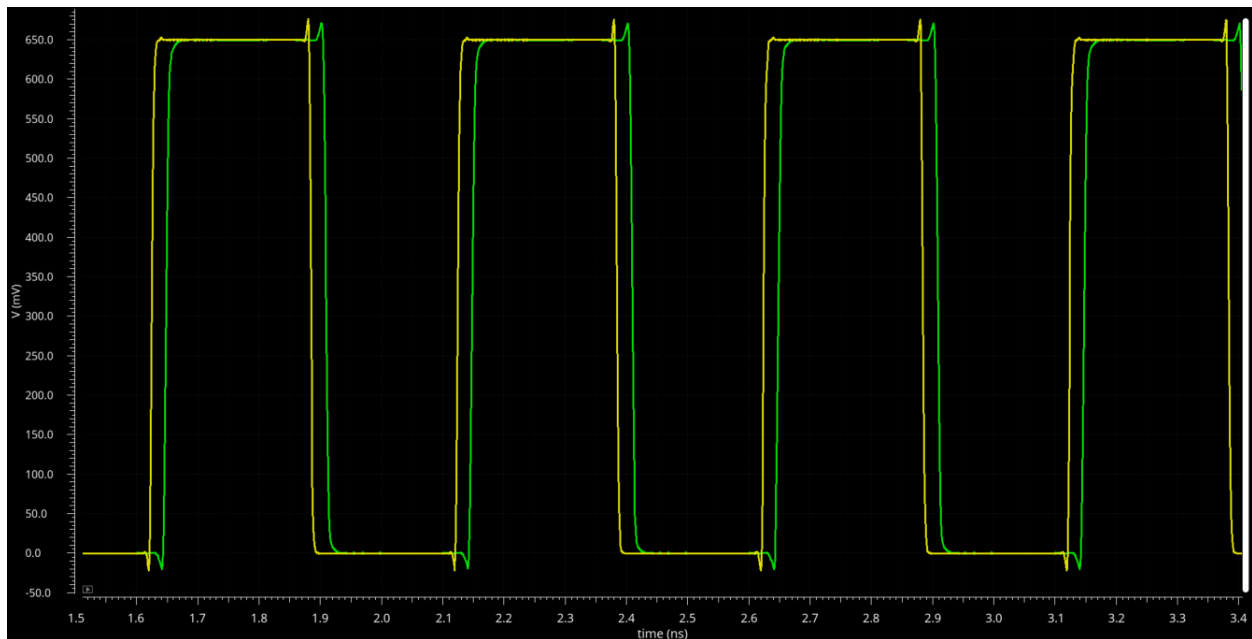
Figure D.3: Full swing VDM's clock (POLO simulations)