Voltage stacking for near/sub-threshold operation

Voltage Stacking for Near/sub-threshold Operation

Kamlesh Singh

Doctorate committee:

prof. dr. J. Pineda de Gyvez	Eindhoven University of Technology, *1e promotor*
prof. dr. H. Corporaal	Eindhoven University of Technology, *2e promotor*
dr. H. Jiao	Eindhoven University of Technology, *copromotor*
prof. dr. P. G. M. Baltus	Eindhoven University of Technology, *chairman*
prof. dr. M. R. Stan	University of Virginia
prof. dr. W. Dehaene	KU Leuven
dr. P. J. A. Harpe	Eindhoven University of Technology

This work is funded by Dutch NWO project 14714 BrainWave.

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-537-47

Voltage Stacking for Near/sub-threshold Operation

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven, op gezag van de
rector magnificus prof. dr. ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op donderdag 21 October 2021 om 16:00 uur

door

Kamlesh Singh

geboren te Bihar, India.

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. P. G. M. Baltus
1^e promotor:	prof. dr. J. Pineda de Gyvez
2^e promotor:	prof. dr. H. Corporaal
copromotor:	dr. H. Jiao
leden:	prof. dr. M. R. Stan (University of Virginia)
	prof. dr. W. Dehaene (KU Leuven)
	dr. P. J. A. Harpe

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Dedicated to the memory of my mother
Kawshalaya Devi

Summary

In recent years, an explosive growth of small, battery-powered devices for low power embedded applications, such as Internet-of-Things, wearable sensors, and biomedical applications are gaining popularity. The convenience of these devices grows with better functionality, processing power, and more importantly battery life. Therefore, one of the focus trends in developing digital circuits and systems has shifted towards energy efficiency. Needless to say, long battery life is relevant especially when periodical battery replacement is impractical, expensive, or impossible, or in always-on devices where minimum energy consideration comes into play.

Normally, to address these challenges, the energy consumption of digital CMOS circuits is reduced by decreasing the supply voltage. Scaling the voltage to the near/sub-threshold region is a convincing technique to achieve low power, but at the expense of performance loss. Yet, scaling the voltage to the near/sub-threshold region comes with many challenges. Firstly, it leads to significant performance degradation which requires innovations at the circuit as well as architecture level. Secondly, process variations limit the design robustness in sub-threshold; a fact that is certainly preconditioning the broad adoption of very low voltage operation as an industry standard. And thirdly, the design enablement infrastructure like standard cell libraries, memories, and IPs, is not always offered for low voltage circuits. In fact, in today's developments, the foundry standard cell library is normally designed for super-threshold operation. Hence, the timing and power characteristics of such standard cell libraries in the near/sub-threshold regime are not optimal. Finally, the foundry-provided conventional high-density SRAMs have limited voltage scalability. The standard cell logic can operate down to the near/sub-threshold voltage range, but not the SRAMs. Moreover, since logic and SRAMs are operating at different supply voltages, multiple voltage sources are required. In the current state-of-the-art, these voltage sources are generated through on-chip voltage converters, which

come at the expense of an obvious overhead of power and area.

The research presented in this thesis advances the state-of-the-art of low-power design in various aspects. From an enablement perspective, design techniques to improve performance and reduce power consumption by using latches are introduced along with a standard cell library pruning technique for near/sub-threshold operation. The selection procedure explores the relative variability in the delay of standard cells when the voltage is scaled to the near/sub-threshold region. New standard cells are introduced as well to minimize leakage and leakage spread in general. But, the main contribution of this research is a new implementation philosophy which is embodied in a three-level voltage stacked system. This design strategy places the foundry-provided high-density SRAMs on the top stack, and standard-cell-based logic, working in the near/sub-threshold regime, in the middle and bottom stacks. Such a design strategy overcomes the limitations imposed by the voltage difference between SRAMs and logic, and it avoids the use of on-chip bulky power converters. The proposed approach has been demonstrated through a RISC-V-based microcontroller which was implemented in a 28-nm FDSOI technology. The proposed voltage stacked system achieves a system efficiency (power required by the system/power supplied to the system) of up to 95%. Both a current sink voltage controller and an adaptive body bias (ABB) voltage controller are used to regulating the intermediate voltage rails between the stacks. This "converter-less" design approach results in significantly large area savings as compared to the state-of-the-art flat and voltage stacked systems. In actual fact, this is the first work in the scientific community where a converter-less voltage stacking scheme is implemented for near/sub-threshold operation achieving 95% system efficiency and that is operating from a 1.7 V (1.8 V $\pm$ 5%) supply voltage.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, due to rapid technological innovations, an explosive growth of small, intelligent, ultra-low-power electronic devices has taken over human lives and the surrounding environment. Many emerging technologies for consumer electronics, such as smart home appliances, wireless sensor networks, gaming devices, audio devices, laptops, desktops, portable battery-operated devices in modern applications such as wearables, fitness trackers, smart devices (watch, glass, phones, etc.), healthcare (biomedical sensors), Internet-of-Things (IoT) devices, etc. require electronic circuits that can function with minimum energy use. These modern devices incorporate a microcontroller-based system as the main computing component that has a significant impact on the power consumption. The market for the ultra-low-power microcontroller is projected to grow with the growth of the consumer electronic market. A compound annual growth of 24.1% is expected in the near future for the ultra-low-power micro-controller market, which will reach up to USD 12.9 billion by 2024 [1].

The battery-powered applications where the majority of the energy consumption is for the digital processing and can operate at a relatively relaxed frequency constraint are suited for ultra-low energy operation. The critical parameters for these devices are run-time, battery lifetime, footprint size, and battery size/volume. For example, wireless sensor nodes, biomedical implants, mobile personal healthcare monitoring, and hearing aids are applications that have a very tight energy budget and therefore should have ultra-low energy consumption. Especially, in biomedical applications to monitor human body parameters, the devices should be light and small to make them comfortable to the body, which causes low battery capacity. Additionally, in some cases, periodical battery refill is impractical, expensive, or impossible. Therefore, reducing the power consumption of these devices enables a long battery lifetime, device portability, and lightweight with reasonable small-size packaging. In the past years, the scientific community has devoted significant research efforts to aggressively reducing the energy consumption of these devices.

This thesis has been done under the umbrella of the **BrainWave** project [2]. The BrainWave project focuses on electroencephalogram (EEG) signal processing for epilepsy and freezing-of-gait (FoG) in Parkinson's Disease. The BrainWave platform targets 24-hour/7-day always-on continuous operation. In this chapter, we first discuss the objectives of the BrainWave project. Then, an overview of the past twenty years of ultra-low-power digital circuit design techniques which motivate the problem statement is presented. Following that, we state the scientific problems that this thesis addresses. Finally, the goal and the organization of this thesis are presented.

1.1 BrainWave project

Brain-related diseases, such as epilepsy and Parkinson's Disease, are life-threatening and severely degrade people's quality of life. Approximately 65 million people worldwide suffer from epilepsy or Parkinson's Disease, making them one of the most common neurological diseases globally [3]. Each year, more than 1 in 1,000 people with epilepsy die [4]. There are approximately 120,000 epilepsy patients in the Netherlands [5]. Annually, the number of newly registered patients in the Netherlands is between 5,000 and 8,000 [5]. Nowadays, EEG-based signal-processing techniques are used for diagnosing and monitoring epilepsy which reflects the electrical activities or disorders of neurons in the human brain. Many patients have to go to specialized hospitals to receive continuous monitoring of their EEG signals, which is costly and impacts the patient's well-being. Furthermore, significant medical attention is required for epilepsy and Parkinson's Disease patients. To monitor the patients, typically, the patient wears a headset with multiple sensors/electrodes (channels) that are wired through long cables to bulky equipment. Furthermore, there are limited means for remote monitoring, e.g. at the patient's home. As a result, patients cannot be monitored 24-hours/7-days, continuously. The existing diagnosis and treatment methods require long-term in-hospital monitoring, which is costly, time-consuming, and are uncomfortable for the patients. The commercial devices existing for ambulatory EEG monitoring generally support only a few EEG channels (e.g. EEG patch) or have limited battery lifetime (e.g. TMSi Mobita, TMSi SAGA 32/64+ [6], g.Nautilus-PRO [7]), or are insufficient to reliably detect more complex brain-related seizure types [8]. Wearable platforms, which can reliably detect epileptic seizures or FoG, would significantly improve the patient's situation. BrainWave delivers such a wearable, low-power platform enabling 24-hours/7-days healthcare of epilepsy and Parkinson's Disease patients in non-hospital environments. This platform will not only improve the patients' quality of life but also save significant amount of money for treatment purposes.

The existing devices such as TMSi Mobita [6], g.Nautilus-PRO [7], support stream-

ing of EEG signals using Wi-Fi for processing on computer or server, else they are used for recording the signals for offline processing. In [9]–[12], the continuous streaming of EEG signals using Wi-Fi (with no processing on the edge), the radio alone was responsible for up to >60% of the total power consumption. The radio power consumption can be reduced by using a micro-controller or digital signal processor (DSP) on the edge to process the raw data and only transmit the important extracted features or transmit an alarm if a seizure is detected. This reduces the radio power consumption at the cost of increasing the processing power. In order to reduce the total power consumption of the system, the power budget for EEG signal processing is very limited. In [13]–[15], the power consumption for the bio-signal processor for real-time EEG seizure detection (pre-processing, feature extraction, and classification) is $<200\,\mu$W for processing data of up to eight channels. The existing state-of-the-art systems are either optimized for simple epileptic seizure detection or signal processing for a limited number of channels, usually up to eight channels [13]–[15].

Achieving ultra-low power consumption requires innovations in all aspects of the design, i.e. system level, algorithm level, architecture level, and circuit level. To obtain a prolonged battery life (>1 week) without compromising the quality, different components in an EEG monitoring system need to be carefully tuned. At the system level, the state-of-the-art platforms utilize 10-bit to 12-bit analog to digital converters (ADCs), low noise amplifiers, and advanced filtering in the Analog Front-End (AFE) to maximize battery life [10], [14], [16]. For emerging and complex EEG monitoring tasks such as non-convulsive epileptic seizure detection and Parkinson's Disease FoG prediction, research is ongoing on what algorithms and sensors work best. A classical EEG-based epileptic seizure classification pipeline consists of data acquisition, pre-processing, feature extraction, and classification. Usually, the EEG signals are acquired at a sampling rate of 100 Hz from 21 channels [17], [18]. Each EEG recording is segmented by a 2.56-second interval, which is called an epoch. Then each channel is pre-processed using a 1 Hz–45 Hz Butterworth band-pass filter [17], [18]. The pre-processing step suppresses low-frequency movement artifacts and interference from 50 Hz alternating current (AC) mains [18]. For each EEG channel, several features can be calculated. For example, basic time-series-based statistical features such as zero-cross times, signal mean and standard deviation, peak features [19]–[21]. Traditional spectral and time-frequency features such as absolute, relative power of the 5-EEG rhythm bands, discrete wavelet transform, non-linear features including Approximate Entropy, and the Hurst exponent can be calculated [19]–[21]. Based on the resulting feature values, the likelihood of a seizure being present in the current epoch is predicted by a machine learning classifier, for example support-vector machine (SVM), Rust-boost classifier [17], [18], [20], [21]. In [21], a survey and feature importance analysis of 47 common EEG features is conducted. In [18], a

detailed analysis of computational complexity and energy breakdown for the EEG-based seizure detection and classification pipeline in [21] is performed. The analysis confirms that, for a system comprising of AFE, 10-bit to 12-bit ADC, edge processing on a microcontroller, and radio, the processing on the microcontroller consumes >95% of the total system energy consumption. Thus, seizure detection and classification demand an energy-efficient and flexible processing platform. Additionally, in the literature, the biomedical signal processing platforms are commonly designed with multiple processor cores and are typically coupled with hardware accelerators [16], [18], [22]–[24]. A combination of programmable and hard-wired processing elements is needed to ensure the ability to map different algorithms while consuming significantly lower power than fully programmable architectures.

Our target is to implement an ultra-low-power microcontroller platform with efficient signal processing capability within the umbrella of BrainWave.

1.2 Ultra-Low-power digital design

Emerging battery-operated applications demand electronic circuits to have reduced energy consumption, portable size while maintaining computation throughput. This results in a more stringent energy budget for long battery life. There are several techniques for reducing energy consumption (E_{cycle}) of a digital circuit. The energy consumption per cycle of a digital circuit is

$$E_{cycle} = Pt_{cycle} = \alpha CV_{dd}^2 + V_{dd}I_{leakage}t_{cycle}. \tag{1.1}$$

The energy consumption comprises dynamic energy consumption (αCV_{dd}^2) and leakage energy consumption ($V_{dd}I_{leakage}t_{cycle}$). The charging and discharging of load capacitances (C) in digital circuits due to activity (α) results in dynamic energy consumption, whereas the static energy depends on the static current consumption ($I_{leakage} \propto e^{\frac{V_{gs}-V_{th}}{\eta V_t}}$) and the time period ($t_{cycle}$) of the circuit's clock. From the above equation, aggressive voltage scaling down to the near/sub-threshold (V_{th}) region of the transistors improves the energy consumption of the system by orders of magnitude [25]–[29]. In the super-V_{th} region, where $V_{dd} >> V_{th}$, reducing the supply voltage decreases the dynamic energy quadratically. As the voltage scales, not only the dynamic power decreases, the leakage power also reduces. However, in the sub-V_{th} region, where $V_{dd} < V_{th}$, the circuit frequency decreases exponentially ($t_{cycle} \propto \frac{CV_{dd}}{I_{leakage}}$). Therefore, a large amount of energy is lost due to leakage power consumption in the sub-V_{th} region of operation because the leakage current is integrated over a longer clock period. In the following part, a literature review of the state-of-the-art ultra-low-power design approaches mainly focused on near/sub-V_{th} designs over the past 20 years is presented.

1.2.1 Twenty years of near/sub-threshold design

Scaling the voltage to near/sub-V_{th} brings not only quadratic dynamic energy savings but also super-linearly reduced leakage current. The fundamentals of near/sub-V_{th} design have been very well explored [25], [26], [28], [30]. Nevertheless, despite that researchers have made big efforts to develop effective techniques and design styles over the years, still, voltage scaling to near/sub-V_{th} is not common in the industry, with of course some exceptions [31], [32]. The designer aims to overcome these challenges and to develop new techniques to meet energy requirements.

Design improvements broadly fall under circuits and architectural techniques. One of the early known sub-V_{th} processors relying on circuit design techniques was presented in 2004 by Wang *et al.* [33]. The chip was a fast Fourier transform (FFT) processor designed in a 180-nm CMOS process operating at 180 mV. The FFT processor was designed using a modified pass-transistor-based standard logic cell library, custom memory with a latch-based register-file, and multiplexed reads to ensure a robust operation. Alternatively, the two-stage micro-architecture processor in [34] based on shallow pipelines with high FO4 delay per stage to reduce variability is an example of architectural improvements. The design of this processor used a careful standard cell selection and a custom static random-access memory (SRAM) to ensure fully functional operation down to 200 mV to consume only 2.60pJ/inst in a 130-nm CMOS technology. Likewise, the processor of [35] is an example that mitigates the impact of process variations by adopting body biasing and gate sizing techniques for sub-V_{th}. Heretofore, modified standard cell structures and latch/mux based register files as memory were used for operation in sub-V_{th} region resulting in significant design, power, and area overhead. In 2006, Calhoun *et al.* [36] proposed the design of a 10T SRAM operating in sub-V_{th}, improving the signal-to-noise margin degradation as compared to 6T SRAM. Later, Verma *et al.* [37] designed a compact 8T SRAM for low voltage operation. Over the years, several SRAM cells have been proposed such as 8T, 9T, 10T, 11T, and 12T to improve functionality at low voltage by increasing read and/or write static noise margins [38]–[43].

Fig. 1.1 puts together the energy consumption and performance of these chips over the years [34], [35], [44]–[66]. The figure depicts the energy consumption of low power microcontrollers (scalar cores), multi-cores, and DSP cores operating in the near/sub-V_{th} region. The energy consumption of the chip proposed in this work is higher than the system in [59] as our design uses 10× more SRAM memory. To fairly compare these chips, we show the energy per operation (pJ/op) for the designs. Usually, scalar processor cores execute <1 operation per cycle (assumed 1 for the plot) [67]. However, the multi-core and DSP platforms can do multiple operations per cycle.

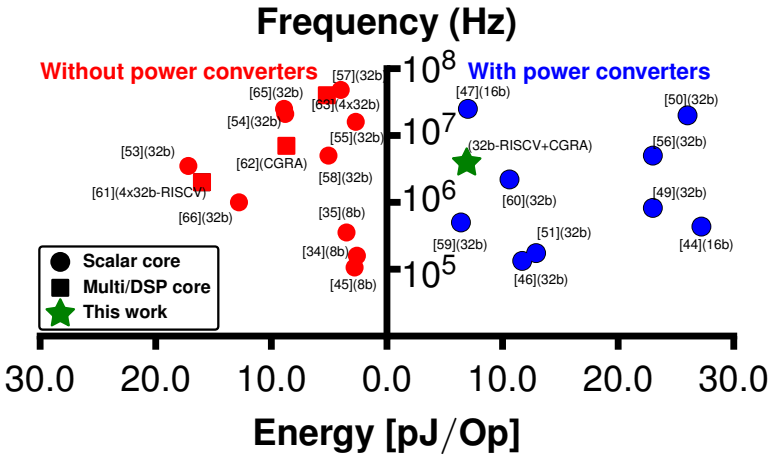


Figure 1.1. Frequency energy correlation of published chips operating in near/sub-V_{th} region. Annotation (CPU bit-width).

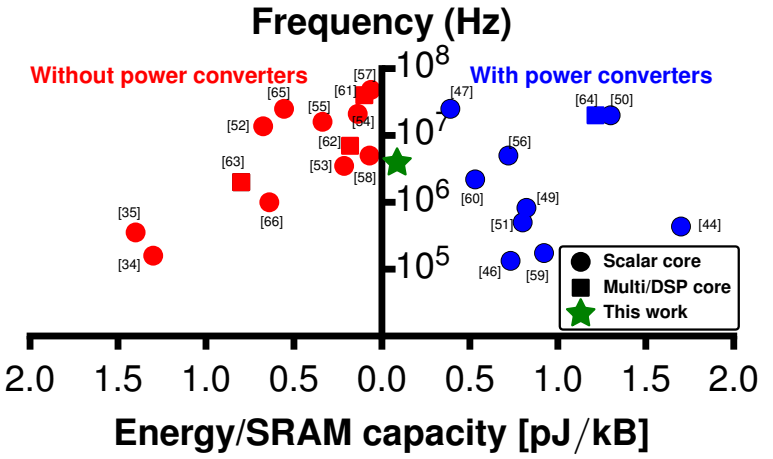


Figure 1.2. Frequency vs. energy consumption per SRAM capacity of published chips operating in near/sub-V_{th} region.

Key factors impacting energy consumption are obviously technology, CPU architecture, operating conditions, power gating, power conversion efficiency, and to a large extent memory type and size. The reported memory sizes go from 256 B up to 150 kB. The larger the active memory size the higher is the energy consumption. To fairly represent the impact of memory size on energy consumption, the performances of the aforementioned designs are plotted against energy per kB of SRAM capacity (pJ/kB, the lower the energy, the better the design is) in Fig. 1.2. Note also that some of the works do not include power delivery losses, which can also be significant and are often ignored. The performance and energy measurement results of the circuit prototype presented in this thesis are depicted in Fig. 1.1 and 1.2. The measured energy consumption per operation and energy consumption per kB of SRAM capacity is competitive to the state-of-the-art. But broadly speaking, these chips represent state-of-the-art architecture and circuit innovations.

Although, voltage scaling to near/sub-V_{th} region results into significant energy gains, it brings along design challenges [50], [68]–[72]. From reviewing these papers, it became evident that the major problems for robust operation of sub-V_{th} designs are process variability and performance degradation. There are several state-of-the-art works to tackle these challenges. Among them, one can find complex chips that use architecture level parallelism [61], [68], [73], specialized multi-core processors [61]–[64], [73], and pipelining techniques [74], [75] to compensate for the throughput degradation while operating in sub-V_{th}. In [68], architecture level parallelism of a JPEG compression co-processor is used to compensate for the throughput degradation. In [73], a very long instruction word (VLIW) processor coarse-grained reconfigurable array (CGRA) with 9 funtional units (FUs) is used for energy efficiency and performance scalability. The system consumes 30pJ/cycle at 400mV in a 40-nm CMOS technology. In [61], a RISC-core specifically designed for near/sub-threshold is used in multi-core clusters. The four core cluster achieves a peak efficiency of 193 MOPS/mW while operating at 40MHz with a power consumption of 1mW in a 28-nm Fully Depleted Silicon on Insulator (FDSOI) technology.

Near-V_{th} computing was also proposed where the supply voltage is approximately set close to the V_{th} of the transistors [26], [28], [76], [77]. Near-V_{th} operation retains much of the energy savings of sub-V_{th} with relatively higher performance, since lowering the voltage further (sub-V_{th} operation) leads to extremely slow transistors. Note that designing circuits for a target performance is difficult due to process variations, hence some tunability is required to achieve performance targets without taking huge design margins.

To illustrate the impact of process and temperature variations on energy and performance we simulated a 99-stage NAND-FO4 based ring oscillator in a 28-nm FDSOI technology. Fig. 1.3 shows the energy spread (normalized w.r.t. TT, 0.9 V,

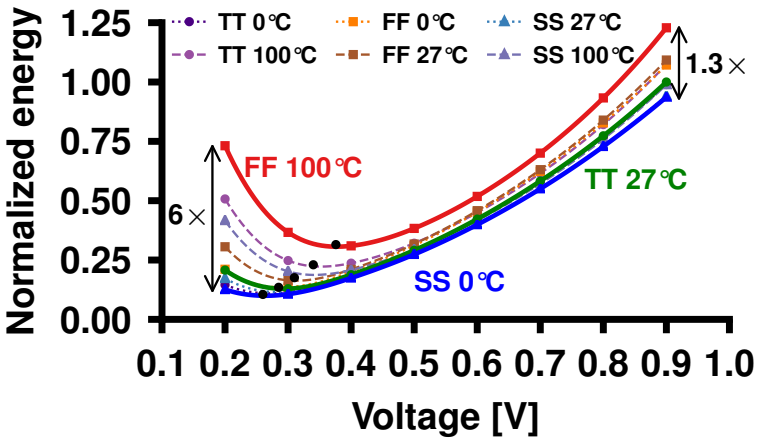


Figure 1.3. Energy vs. voltage for a 99-stage NAND-FO4 based ring oscillator with process and temperature variations.

27°C) with voltage scaling. In this plot, the dynamic energy ($\propto CV_{dd}^2$) doesn't change much with temperature. However, the static energy consumption significantly depends on temperature and V_{th} variations. The energy consumption varies by 6× in sub-V_{th} and 1.3× in super-V_{th} primarily because of temperature change. The energy consumption varies within 2× (SS−FF) with process variation at a given temperature. Delay spread with process and temperature variations is shown in Fig. 1.4. One can see that the delay variation increases from ∼1.4× in super-V_{th} to ∼125× in sub-V_{th} with process and temperature variations. Also, be aware that energy consumption in sub-V_{th} is relatively less sensitive to process and temperature variations as compared to the delay variations as the variations in leakage power consumption and the delay compensate each other.

To mitigate the impact of process variations, several works used adaptive body-biasing (ABB) over the years [35], [50], [57], [58], [68]–[70], [78]–[80]. In the literature, most of these circuits are complex and require a significant design effort. Usually, switched capacitor charge pumps are used to generate the positive and negative bias voltages [50], [69], [71], [72], [79], [80]. In [50], a 32-bit SPARC-V8 processor system-on-chip (SoC) with a closed-loop dynamic compensation for temperature and process variations to enable constant frequency is designed in a 28-nm FDSOI technology. Switched capacitor-based positive and negative voltage generators are used to generate the forward body-biasing (FBB) for NMOS and PMOS circuit. In [80],

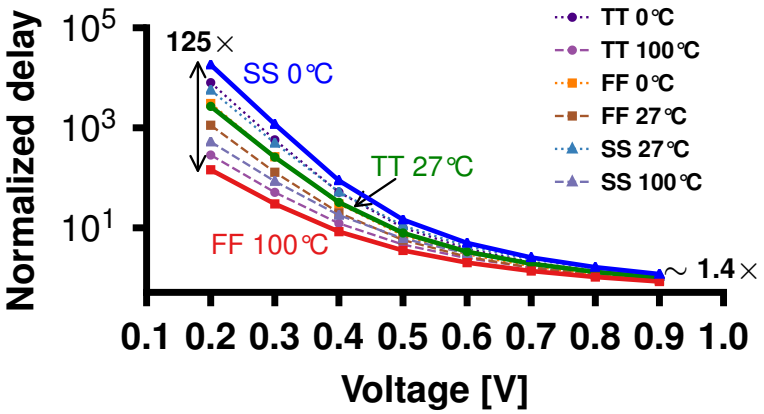


Figure 1.4. Delay vs. voltage for a 99-stage NAND-FO4 based ring oscillator with process and temperature variations.

a body-bias generator is presented with a fine voltage step and less than 100 ns response time in a 28-nm FDSOI technology. The power and area overhead for the body-bias controller are $<10\mu$W and 1.2%, respectively. In [71], an ABB SoC using in-situ timing slack monitoring for a tunable replica circuit is presented. The system achieves up to 53% energy improvement due to reduced margins. Unlike other works, Pu *et al.* [68] proposed the design of a simple V_{th} sensor/actuator to balance the V_{th} voltage mismatch between PMOS and NMOS transistors. The V_{th} balancer is a simple design based on a CMOS inverter, whose PMOS and NMOS transistors are off functioning as a process-corner V_{th} imbalance detector. In contrast to other works, this approach makes use of only one body-controlling line for both PMOS and NMOS.

To dynamically control the body bias voltage various process monitoring techniques have been devised e.g. critical path replica based monitoring, or sets of multiple ring oscillators to determine the process and temperature conditions [58], [69], [71], [79], [81]–[83], and leakage current accordingly [59], [68], [84]. In [85]–[87], shadow latch based timing error detection and correction methods were used to reduce process/temperature/voltage/ageing design margins.

In addition to the above-discussed challenges, power delivery is often ignored. Researchers usually do not consider the energy overhead to generate these ultra-low voltage power supplies. Early works on this subject matter were published in 2008

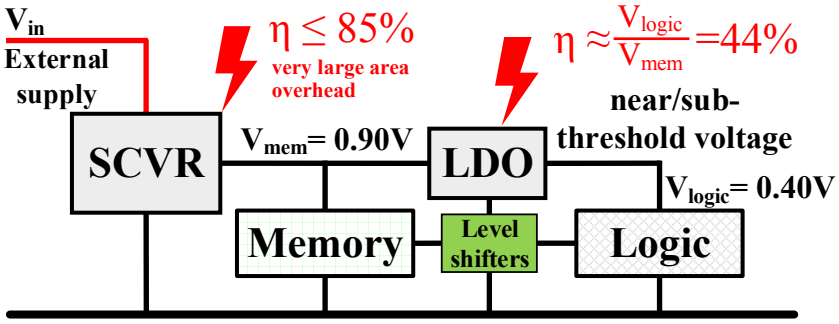


Figure 1.5. Conventional power delivery strategy for low voltage design.

by Kwong *et al.* [44]. They presented a SoC with a 16-bit processor (MSP430) and a 128 kb 8T SRAM operating down to 300 mV, powered by an integrated on-chip DC-DC converter. Additionally, a separate power supply of 1.8 V is often required for industry standard IO-pads. Nowadays, the industry is moving towards 1.2 V as IO-pad supply voltages for battery operated devices.

Another challenge for power delivery is that foundry provided SRAMs have a limited voltage scalability range. Consequently, multiple voltage regulators are needed thus increasing the complexity of the design [47], [49], [88], [89]. Yet, many state-of-the-art designs operating in near/sub-V_{th} use custom designed 8T/10T SRAMs for voltage scaling in sync with the logic circuit [44], [50], [51], [61], [64]. Observe that designing an SRAM for low voltage operation comes with area overhead [57], [90]. Consequently, [47], [49], [51], [57] presented designs with foundry SRAMs operating at nominal supply, the logic operating at sub-V_{th} voltage, and with an integrated power delivery system. In such SoC, the power delivery made use of multiple power domains and multiple voltage regulators as shown in Fig. 1.5. Efficient on-chip switched-capacitor voltage regulators (SCVRs) and low-dropout regulators (LDOs) are mainstream nowadays [49], [91], [92]. However, note that in a more complex SoC, the use of multiple voltage converters bring in extra power consumption and area overhead. Most of the converters in the literature are designed for high payloads, high output voltage, and are not suitable for near/sub-V_{th} [93]. Thus, a fully integrated voltage regulator with a low voltage conversion ratio (V_{out}/V_{in}) and high efficiency is needed. However, a low voltage conversion ratio results in low efficiency for linear regulators and fully integrated buck converters. On the other hand, SCVRs can achieve relatively high efficiency at low voltage conversion ratios [91], [92], [94], [95]. Additionally, reconfigurable SCVR designs with multiple conversion ratios can in fact provide low output voltages while maintaining high efficiency. In [91], four

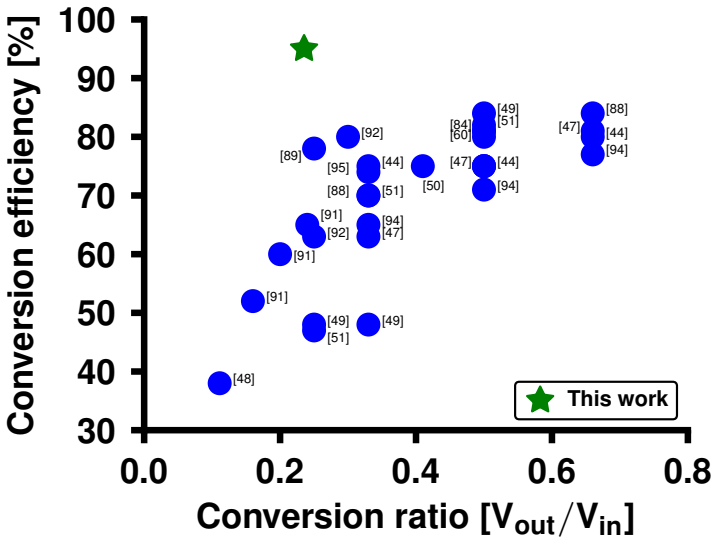


Figure 1.6. Power delivery conversion efficiency vs. voltage conversion ratio.

power stages are cascaded in recursive sequence to achieve 15 linear ratios, providing a low output voltage down to 0.1 V, achieving 55% conversion efficiency for 0.5 V (conversion ratio of 0.2) output voltage from a 2.5 V input supply. However, this requires multiple switches and flying capacitors. In [64], the designed system, operating in the near/sub-V_{th} region, achieves a system efficiency ($1 - P_{loss}/P_{in}$, see equation (5.2)) between 82–89% with conversion ratios of 2/3 and 1/2. The system uses an on-chip SCVR which generates an output voltage of 0.48 V using a 1.0 V (1/2 ratio). It uses two input supplies of 1.0 V and 1.8 V to generate other voltages with 36% area overhead. The 28-nm FDSOI system in [89] operates in the near/sub-V_{th} region at 0.3 V–0.5 V. The near/sub-V_{th} voltage is generated using an on-chip 1/3 DC-DC converter from a 1.55 V lithium-ion battery, achieving a total system efficiency of 80% with an area overhead of 38%. However, [89] does not account for the SRAM supply voltage. In [47], the implemented system operates at 0.4 V generated using an on-chip DC-DC converter from a 1.0 V (1/2 ratio) supply. The overall system achieves an efficiency of up to 80% with an area overhead of ∼30%. Furthermore, this system also requires an additional supply voltage for the IO-cells. In [96], a fully integrated SCVR in 28-nm FDSOI is presented that generates four on-chip voltages of 0.5 V, 0.67 V, 0.9 V, and 1 V using 1 V-core and 1.8 V IO voltage inputs. The

SCVR achieves 80–86% conversion efficiency with 16% area overhead. Note that the designed system still requires two distinct supply voltages for IO-cells and core, resulting in a higher overall off-chip system overhead. In [92], multiple SCVRs are designed for the near/sub-V_{th} region in a 65-nm CMOS technology. The designed SCVRs achieve a conversion efficiency of up to 80%. In general, the available on-chip voltage regulation schemes for output voltage in the near/sub-V_{th} region achieve system efficiencies up to 85% [44], [47], [49], [57], [59], [64], [84], [88], [89], [92], [94]–[97]. Fig. 1.6 represents the state-of-the-art achieved power delivery efficiency vs. the conversion ratio to generate near/sub-V_{th} supply voltages.

Given the knowledge of the challenges from the literature review, this doctoral research is formulated in the next section.

1.3 Motivation and Problem statement

The review of the past twenty years of literature highlights the interest of the researchers and improvements made in the respective directions. From these state-of-the-art works, it is clear that simply scaling the supply voltage without consideration of its effects would not result in optimal energy operation. Looking at the state-of-the-arts, some key considerations important for this thesis in regards to near/sub-V_{th} design are as follows:

- **Performance degradation:** As stated in the previous section, performance degradation is a big drawback for the near/sub-V_{th} designs. At large, the performance degradation is addressed by parallelism and/or pipelining [61]–[64], [68], [73]–[76]. Besides, the performance degradation results in several interesting architecture design opportunities for memory and logic core subsystem [61], [73], [76].

- **Process and temperature variations in near/sub-V_{th} design:** Process and temperature variations are the biggest hurdles in the industrialization of designs operating in the near/sub-V_{th} region. In the digital design flow, simply guaranteeing performance under the worst-case conditions is not enough for the near/sub-V_{th} operation. Furthermore, the margins in the near/sub-V_{th} design can grow to the extent that they neutralize the energy gains when all different process and temperature variations are considered in the design flow. There are several techniques to mitigate the impact of process and temperature variations. These techniques extend from proper sizing of transistors in the standard library cells to ABB of the design.

- **SRAMs and near/sub-V_{th} design:** Voltage scaling of the logic circuit is not the same as that of SRAM blocks. Therefore, either a special custom designed low voltage SRAM is required for which the supply voltage scales in sync with the logic circuit, or the design is partitioned in multiple voltage domains separating the foundry provided SRAM and logic circuits. Furthermore, the custom design low voltage SRAM requires significant design and characterization effort to enable an integrated digital design. Additionally, the designed low voltage SRAMs have significant area overhead ($\sim 6\times$ as compared to foundry provided SRAM) [57], [90].

- **Integrated power delivery for near/sub-V_{th} operation:** As the complexity of a SoC is increasing (multiple cores, power domains, etc.), there is an increasing requirement for using integrated power conversion for the chip in order to optimize the total system power and performance. For example, in Fig. 1.5, for a near/sub-V_{th} design, a straightforward choice of using the foundry provided high-density SRAMs results in the use of multiple voltage regulators in the system. This choice results in a significant amount of energy that is consumed by the power conversion losses. In [51], up to 40% of the total energy consumption is dissipated by the integrated SCVR, while operating in the near/sub-V_{th} region. The use of multiple supplies for the power domains significantly reduces the overall system efficiency. The overall gains achieved due to voltage scaling diminish. These power delivery networks are designed in a hierarchical manner, combining slower and more efficient SCVRs with faster and less efficient LDOs. Additionally, the requirement of multiple supply voltage also increases the design and area overhead.

The low power circuit design level techniques, process variation mitigation techniques, and system architecture level optimizations are well explored in the literature. However, the developments in the near/sub-V_{th} operation with embedded power delivery techniques still has room for improvements. The power delivery for low voltage operation have special requirements over existing power converters. In the state-of-the-art, usually, the power delivery circuitry (SCVR, LDO) is over-designed to meet the worst condition and the conversion efficiency is significantly low for relatively low current loads. Therefore, the designed circuits might behave optimally standalone; but when integrated into the complete system, they might have sub-optimal behavior. Below we present the design challenges and complexity related to the SCVR and LDO for near/sub-V_{th} power delivery.

- **SCVR design:** The fully integrated SCVR utilizes capacitors and power switches available in the CMOS process. The number of required capacitors and power switches depends on the conversion ratio. For example, typically, for a

divide-by-2 SCVR, at least one capacitor (C_{fly}) and four switches are required which operate in two phases, each of which ideally has 50% duty cycle (switching frequency) [49], [95], [98]. The capacitor size and switching frequency are decided based on the power density of the load. Furthermore, the voltage ripple at the output of the converter has a direct impact on the conversion efficiency. The voltage ripple depends on the capacitor size and switching clock duty cycle. Additionally, the SCVR are only efficient at discrete conversion ratios (1/2, 2/3, 1/3, 1/4, etc.) of input-to-output voltage [51], [64], [89], [98]. Besides that, the conversion efficiency of the SCVR is highly dependent on the Q-factor of the capacitors and losses across the switches [88], [95], [98]. Moreover, often, the conversion ratio doesn't lead to the expected output voltage, therefore additional closed-loop control circuitry is required to stabilize the supply at the desired voltage [89]. The design and trade-off of SCVR are well studied in the literature [49], [51], [64], [88], [89], [93], [95], [98].

- **LDO design:** Low-dropout linear regulators are usually used to convert noisy supply voltages to a low noise accurate voltage. The main components of an LDO are a pass-transistor power MOS, an error amplifier, and a voltage reference. Traditionally, LDOs have been analog in nature and employ a high-gain error amplifier to control the pass transistor output and provide regulation. Although LDOs are small and achieve fast response times [99], they are highly inefficient ($\propto V_{out}/V_{in}$), potentially limiting the system efficiency [49], [100]. Usually, an LDO is used in cascade to an SCVR to reduce the voltage supply noise. However, the analog design nature of traditional LDOs does not allow operation at low supply, and low control voltages make it difficult to integrate for near/sub-V_{th} output voltage, resulting in the design of digital LDOs [101].

Looking at the existing challenges, the problem statement for this thesis is: ***How to enable design techniques to realize a robust and ultra-low-energy system with integrated power delivery operating in the near/sub-V_{th} region with extremely low power conversion losses, less complexity of custom design, and less area overhead.*** The discussed design challenges of the system operating in the near/sub-V_{th} region are tackled at four levels of the design: the system level, the architecture level, the circuit/gate level, and the physical design level.

1.4 Goal and overview of this thesis

The previous sections of this chapter introduced the state-of-the-art near/sub-V_{th} designs and the most important challenges. The focus of this thesis is primarily pushing the state-of-the-art to provide techniques and methodologies at different

abstraction levels of design to enable a robust, energy-efficient, and ultra-low-power system operating in the near/sub-V_{th} region. Additionally, the thesis is motivated towards an integrated design approach for the digital integrated circuits (ICs) with a focus on new opportunities for an embedded power delivery for near/sub-V_{th} designs. The key challenges discussed in previous sections are addressed in this thesis.

The systems operating in near/sub-V_{th} region suffer from a degraded performance which is mitigated by pipelining or parallelism. In Chapter 2, we explore latch-based design as a possible option to enhance the circuit performance. Typically, digital ICs are designed using flip-flops as sequential elements in the pipeline. The used flip-flops can be replaced by latches to improve performance. In a pipeline, the latches provide the flexibility of distributing the timing budget between neighboring stages such that time borrowing is possible [102]–[106]. Usually, latch-based pipelines are used in high performance processors [104]. In the literature, the latch-based design is mostly explored for custom-designed pipeline stages such as in finite impulse response (FIR) filter, shift-register, and multiply-accumulate (MAC) units [107], [108]. Nevertheless, an effort to convert processors to latch-based design is also made in [109], [110]. However, the latch-based processors consumed more power as compared to the original flip-flop based designs. In Chapter 2, we explore a new automatic design flow of converting any flip-flop based design to a latch-based design. Based on the proposed smart retiming technique an optimal operating point is identified for achieving the maximum performance improvements. The flow is initially established for super-threshold operation and later extended to near/sub-V_{th} operation.

The problem related to the variability of the near/sub-V_{th} operation is addressed in Chapter 3. The commercially-available foundry provided standard cell libraries are optimized for super-V_{th} operation. Often, not all cells have a robust operation in the near/sub-V_{th} region of operation. In Chapter 3, we design a process variation aware standard cell library for near/sub-V_{th} operation in a 28-nm FDSOI technology. The main achievement is balancing the worst rise and fall delays without changing the height of the cells, poly separation, area, while still being compatible with the existing standard cell library. Additionally, in Chapter 3, we propose a new thorough pruning methodology of foundry provided commercial standard-cell libraries for near/sub-V_{th} operation. The commercial foundry-provided standard cell libraries can meet the functional yield constraint up to a certain voltage limit in the sub-V_{th} region. Therefore, in the literature, several works propose generic guidelines for standard cell library pruning for near/sub-V_{th} operation [27], [34], [47], [68]. For the first time, we take a quantitative approach towards a rating (*degradation factor*) of a standard cell relative to other cells for its behavior for voltage scaling from super-V_{th} to near/sub-V_{th} region. Based on the relative behavior we propose a library pruning method to filter cells that are not robust in the near/sub-V_{th} region in Chapter 3.

In Chapter 4, we investigate opportunities for energy savings for near/sub-V_{th} operation from a power delivery perspective. The power delivery losses are a significant contributor (up-to ∼40%) to the total energy consumption for a system operating in the near/sub-V_{th} region. Therefore, a new perspective is necessary for the power delivery challenges in the near/sub-V_{th} region to save energy. Our approach to looking towards other techniques instead of existing conventional methods, as shown in Fig. 1.5. We investigate voltage stacking as a possible prospect for integrated power delivery for the near/sub-V_{th} operation. Voltage stacking is based on Kirchhoff's voltage law for series-connected power domains such that the ground of one domain becomes the power connection for the next. Thus, the domains are connected in a series stack for power delivery with all of them sharing the same current, and hence the charge is recycled [111]. In the near/sub-V_{th} region, the leakage power consumption is the dominant power source. Voltage stacking recycles the waste leakage current to perform a useful task. In Chapter 4, we present a converter-less voltage stacking power delivery that supplies all of IO-pads, cores, and foundry-provided SRAMs using a single external voltage supply. To the best of our knowledge, this is the first work where a converter-free voltage stacking scheme is implemented for near/sub-V_{th} operation.

In Chapter 5, we present the silicon measurement result of the designed prototype demonstrating the feasibility and gains of voltage stacking for near/sub-V_{th} operation. Additionally, the implemented voltage stacking technique discussed in Chapter 4, deals with several challenges related to the near/sub-V_{th} designs. Firstly, due to the limited voltage scalability of foundry provided SRAMs, the majority of the state-of-the-art systems operating in the near/sub-V_{th} region either require custom designed SRAMs that can scale voltage in sync with the logic or require an additional power supply. The custom design SRAMs leads to significant area overhead as well as significant design and characterization time overhead. In the presented voltage stacked system, the leakage current of the foundry provided SRAMs on the top stack is recycled to sustain the operation of the logic stacks in the near/sub-V_{th} region. Secondly, voltage stacking inherently shows resilience to the process and temperature variations [112]. The gains of voltage stacking for the process and temperature variations are demonstrated in Chapter 5. Finally, in Chapter 5, we demonstrate the advantages of parallelism to mitigate the performance degradation because of voltage scaling to near/sub-V_{th}.

Overall, the design techniques discussed above require a holistic and concurrent design approach at the different abstraction levels of design. The important techniques explored in this thesis are shown in Fig. 1.7 as per their abstraction level. Every abstraction level impacts the energy consumption of a design. The lowest energy consumption can only be achieved when effective design choices are made

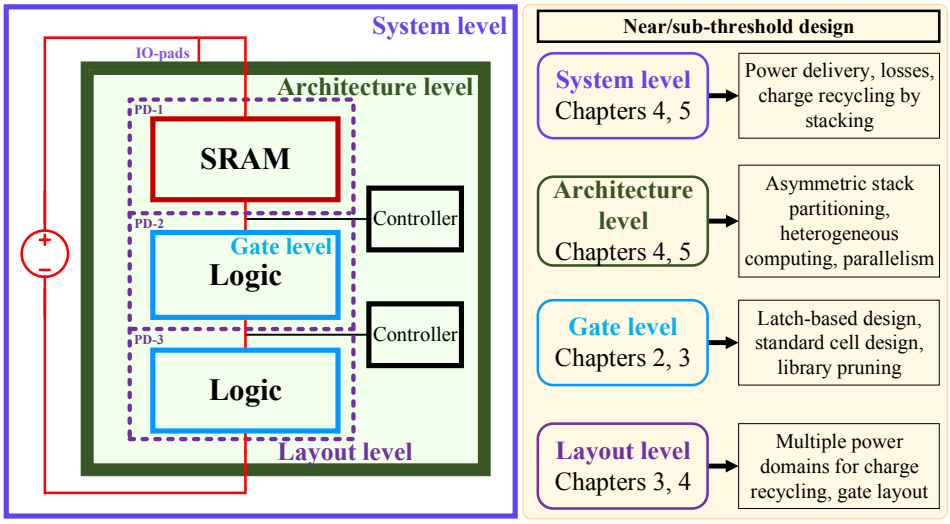


Figure 1.7. Thesis organization at different abstraction levels.

at each abstraction level. In summary, at the system level, the power delivery aspects for a system operating in the near/sub-V_{th} region is explored with the focus on improving the total system efficiency by reducing the number of supplies, power conversion losses, and overall system design complexity. To address these requirements charge recycling by voltage stacking for near/sub-V_{th} is proposed. A three-level voltage-stacked system with SRAMs on the top stack and standard-cell-based logic in the middle and bottom stacks is implemented, as shown in Fig. 1.7. At the architecture level, the proposed voltage stacking is enabled by an asymmetric voltage division between the foundry-provided SRAMs operating in the super-V_{th} region and the logic circuits operating in the near/sub-V_{th} region. Additionally, a heterogeneous parallel compute accelerator is used to compensate for the reduced throughput due to voltage scaling to near/sub-V_{th} region. At the gate level, a tool flow is proposed to convert any flip-flop based design to latch-based design to mitigate the performance degradation in near/sub-V_{th} operation. Additionally, the design of low power robust standard cell library and pruning of existing library for reliable near/sub-V_{th} operation are investigated at the gate level. Finally, at the layout level, multiple voltage domains, body-bias islands, and charge recycling by voltage stacking techniques are implemented. Through this holistic approach of optimizing at different abstraction levels significant power and energy savings are achieved in this work.

Chapter 2

Latch-based Design

Portable/wearable devices and IoT applications are ubiquitous nowadays. As the heart of those devices, intelligent ICs play a pivotal role in the applications. ICs in mobile or IoT applications are either powered by a battery with limited volume or scavenge energy from the surrounding environment. The power/energy consumption requirement for those ICs is therefore rigid. Usually, to meet the required power/energy budget the supply voltage is scaled down to the near/sub-V_{th} region. Voltage scaling to the near/sub-V_{th} region decreases the energy consumption, allowing the circuit to achieve minimum energy per operation. However, due to exponentially large delays of circuits in the near/sub-V_{th} region, the frequency of operation is decreased.

In the literature, several techniques are used to boost the performance of the circuit such as pipelining, parallelism, etc. [113]. A pipelined circuit has a small combinational delay in each pipeline stage and therefore can have substantially higher clock frequency. In parallelism, the logic circuit is duplicated to allow computation in parallel thereby increasing the circuit throughput. In addition to the architectural level design overhead, these techniques may result in the excessive sequential design, clock-power, and area overheads. In digital design, synchronous circuits are typically implemented using edge-sensitive flip-flops. An important feature of a flip-flop-based digital circuit is that the maximum achievable operating frequency of the circuit depends on the propagation delay of the longest path in the pipeline stage. Since flip-flops present such hard boundaries between pipeline stages and if one stage compute in less time, this slack cannot be passed onto other stages to allow longer computation time. Therefore, the flip-flop-based design method is a worst-case design method. This feature also represents an important drawback of flip-flop-based design, especially in high-performance circuits where clock skew and jitter tend to dominate the clock cycle [103]. The advantage of the flip-flop-based design is its resilience against duty cycle jitter. Additionally, the design, verification, and test of digital circuits that are designed using flip-flops are well-supported by commercial

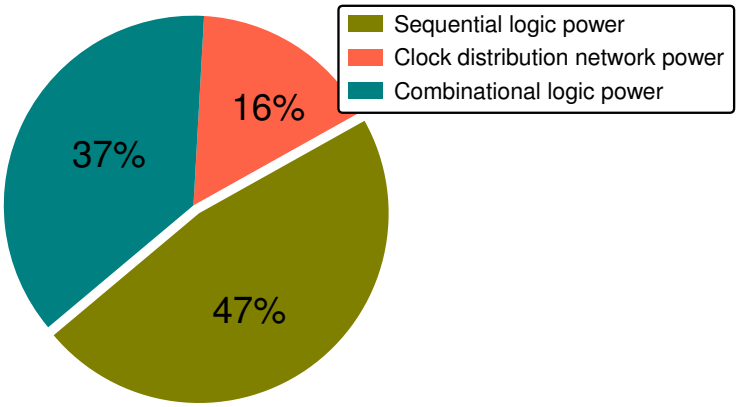


Figure 2.1. Power consumption breakdown for ARM-Cortex-M0.

electronic design automation (EDA) tools. In digital circuits, the flip-flops, as well as the driving clock distribution network, account for up to ∼70% of the total power consumption in high-performance as well as ultra-low power ICs [102], [114]–[116]. The power breakdown of an ARM core synthesized using a 28-nm FDSOI shows that the sequential circuit and the clock distribution network consume 63% of the total power consumption, as shown in Fig. 2.1.

Alternatively, latches are seldomly used in digital design. Latches are smaller, faster, and more energy-efficient as compared to flip-flops. Latch-based design can eliminate the cycle time overhead of clock uncertainty and skew by balancing the circuit delays between latches [102]–[106]. Latches allow designers to exploit clock skew scheduling to improve cycle time. Designers use latches mainly to reduce the sequencing overhead in high-performance processors [104]. Latches provide flexibility of distributing timing budget between neighboring pipelined stages such that time borrowing is possible [102]–[106], thereby further enhancing the speed of the circuit. Additionally, latches are transparent during the active clock period, allowing time borrowing which result in a design resilient to process variations, especially, for the circuits operating in the near/sub-V_{th} region [75], [110]. Latch-based designs, however, have longer hold time requirements compared to flip-flop-based designs. In short-path-dominant designs, the excessive buffers that are used to fix the hold time violations results in significant power overhead [107].

There are a few investigations of latch-based design in the literature. In earlier

work, the latch-based design is either custom designed or simple circuit blocks (FIR, MAC) are used [107], [108]. In high-performance microprocessors (Alpha 21164) the critical path is custom designed using latches to reduce the timing overhead and enhance the performance [117]. In [107], flip-flop-based designs of FIR filter, shift register, and MAC unit are converted to latch-based designs. Up to 45% of energy savings is achieved by the latch-based designs as compared to flip-flop-based designs in sub-V_{th} region. Similar work is introduced in [108] with the implementation of the latch-based FIR filter. Compared to the conventional flip-flop-based filter, this latch-based filter reduces energy consumption by more than 25%. In [109], an ARM Cortex-M3 is converted to a latch-based design to eliminate the timing margins by using Bubble Razor, which unfortunately consumes more power compared to the original flip-flop-based design. In [110], a latch-based 32-bit icyflex2 processor is implemented, showing minimum energy consumption per operation as low as 17.1 pJ/cycle at 19 kHz and near/sub-V_{th} supply voltage of 0.37 V. The existing work for latch-based design focuses on the frequency improvements, however, lacks the in-depth analysis and trade-off of converting a flip-flop-based design to a latch-based design concerning power/energy consumption, clock tree distribution overhead, and area overhead.

In this chapter, latch-based design is explored as an alternative to flip-flop-based sequential design to improve the performance and power/energy consumption for digital circuits operating in the super-V_{th} region as well as in the near/sub-V_{th} region. The timing and power analysis for both flip-flop-based and latch-based designs are performed. The trade-offs for converting a flip-flop-based design to a latch-based design are formulated. The design flows of converting a flip-flop-based design to a latch-based design as well as a latch/flip-flop-mixed design are proposed. Based on a smart retiming strategy, the optimum operating condition for the latch-based design is identified for achieving the maximum time borrowing, and hence the highest power savings by scaling supply voltage. Finally, the proposed flow is evaluated for the near/sub-V_{th} region of operation.

2.1 Timing and power analysis

Power dissipation in digital CMOS circuits has two major components: dynamic switching power consumption and leakage power consumption. The dynamic power consumption is the major component in designs operating in the super-V_{th} voltage region, while leakage power consumption plays a critical role in designs operating in the idle mode for most of the time or in near/sub-V_{th} region. Without loss of generality, we ignore the power dissipation by short circuit current in the analysis.

The power dissipation of a digital circuit is

$$P = P_{dynamic} + P_{leakage}, \tag{2.1}$$
$$= \alpha C_L V_{dd}^2 f + I_{leakage} V_{dd}. \tag{2.2}$$

$P_{dynamic}$ is the dynamic power consumption, where C_L is the loading capacitance, f is the clock frequency, and α is the activity factor. $P_{leakage}$ is the leakage power consumption. $I_{leakage}$ is the leakage current which consists of sub-V_{th}, gate, and substrate junction leakage currents. The timing-driven power analysis of flip-flop-based and latch-based designs is performed in this section for super-V_{th} region for simplicity. Therefore, leakage power consumption is ignored for simplicity and more attention is paid to the dynamic power consumption.

The flip-flop-based and latch-based pipeline structures are shown in Fig. 2.2. In this section, we try to relate the time borrowing property in latch-based design to the power consumption of the circuit. According to Fig. 2.2a, the timing constraint for the flip-flop-based design in terms of equivalent logical depth (assuming L_{DF2} is the critical path), setup time (T_{SU}), clock skew (T_{SKEW}), and clock period (T_{CLKF}) is

$$T_{CLKF} \geqslant T_{CQ} + \tau_g L_{DF2} + T_{SU} - T_{SKEW}, \tag{2.3}$$

where T_{CQ} is the clock-to-Q propagation delay. τ_g is the equivalent single gate delay. The maximum operating frequency is decided by the timing critical path. The total delay from input to output requires $3 \times T_{CLKF}$, as shown in Fig. 2.2a. Even if the logic depth between the pipelines are not the same, the total delay required is decided by the critical path in the stages.

The latch-based design has twice the number of flip-flops in the design. The combinational logic in the pipeline stages of the flip-flop-based design is divided into the latch pipeline stages. The condition for latch based design to be faster than that of the flip-flop-based design when the combinational logic in the pipelines are not well-balanced. In a complex processor system, it is highly possible that the pipelines are unbalanced (i.e. $\tau_g L_{DF1} \neq \tau_g L_{DF2} \neq \tau_g L_{DF3}$).

For the latch-based design shown in Fig. 2.2b, where each flip-flop is split into one positive latch and one negative latch, the clock period (T_{CLKL}) can be written in terms of logical depths ($L_{DL1}+L_{DL2}+L_{DL3}+L_{DL4}+L_{DL5} = L_{DL} = L_{DF1}+L_{DF2}+$

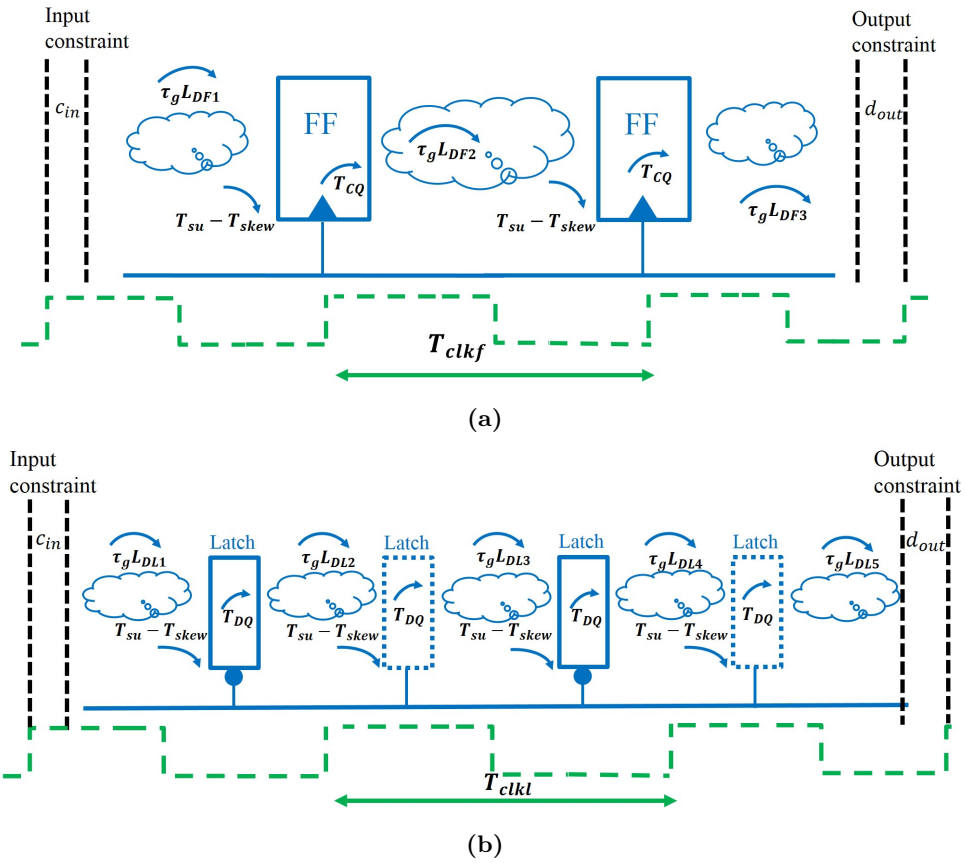


Figure 2.2. Flip-flop-based and latch-based pipeline structures.

L_{DF3}), assuming that the inputs to the latches arrive when they are transparent.

$$\begin{aligned} 3T_{CLKL} \geqslant\ & c_{in} + \tau_g L_{DL1} + T_{DQ} + \tau_g L_{DL2} + T_{DQ} \\ & + \tau_g L_{DL3} + T_{DQ} + \tau_g L_{DL4} + T_{DQ} + \tau_g L_{DL5} \\ & + d_{out} - T_{SKEW}. \end{aligned} \tag{2.4}$$

$$T_{CLKL} \geqslant \frac{c_{in} + \tau_g L_{DL} + 4T_{DQL} + d_{out} - T_{SKEW}}{3}, \tag{2.5}$$

where T_{DQ} ($= T_{DQ+} = T_{DQ-}$) is the D-to-Q propagation delay of a latch. Eq. 2.5 shows the minimum clock period achieved by a latch-based design. It is clear from the above equations that the T_{CLKL} for latch-based is smaller than compared to the flip-flop-based design ($L_{DL1}+L_{DL2}+L_{DL3}+L_{DL4}+L_{DL5} = L_{DL} = L_{DF1}+L_{DF2}+L_{DF3}$

and $L_{DL} < 3 \times L_{DF2}$). There are other timing conditions possible for the latch-based design with various transparent and opaque conditions of the latches. For latch-based design, because of time borrowing (data flow from one stage to the next stage when the latches are open), different pipeline stages can share the available timing slack. Therefore, latch-based designs have the chance to operate at higher frequencies ($T_{CLKF} > T_{CLKL}$) compared to a flip-flop-based designs. Additionally. the latch-based design doesn't have the timing overheads related to the latch (setup time). From (2.5), the latch-based design for any pipe-lined circuit displays advantage over flip-flop-based design in terms of skew and jitter tolerance.

The power consumption of a flip-flop-based design is

$$P_{dynamic-f} = \frac{\alpha_f C_f V_{dd}^2}{\tau_g L_{DF2} + T_{CQ} + T_{SU} - T_{SKEW}}. \tag{2.6}$$

For pipeline circuits, the time borrowing can be accumulative from the first stage to the last stage, thereby resulting in shorter clock period. The power consumption of a latch-based design is

$$P_{dynamic-l} = \frac{\alpha_l C_l V_{dd}^2}{\frac{c_{in}+\tau_g L_{DL}+4T_{DQ}+d_{out}-T_{SKEW}}{3}}. \tag{2.7}$$

The trade-off of converting a flip-flop-based design to a latch-based design can be conceived from (2.13). The factors effecting the power consumption in a latch-based design are activity factor and load capacitance, which change during the conversion.

2.1.1 Power and energy consumption analysis

In the super-V_{th} region of operation, the dynamic power/energy consumption dominates the total power/energy consumption. Therefore, we can scale the supply voltage for the latch-based design to save energy while operating at the same frequency as of the flip-flop-based design. For accessing the scaled voltage, we can express the critical path delay in terms of logical depth and equivalent gate delay. The flip-flop/latch delay, setup time, and skew can be modeled in terms of a certain number of equivalent gate delays. So we can write $\tau_g L_{DF2} + T_{CQ} + T_{SU} - T_{SKEW} = N_f \tau_g$ for flip-flop-based and $\frac{c_{in}+\tau_g L_{DL}+4T_{DQ}+d_{out}-T_{SKEW}}{3} = N_l \tau_g$ for latch-based designs. Therefore, the power consumption of the flip-flop-based and latch-based designs ignoring the leakage power consumption can be rewritten as

$$P_{dynamic-f} = \frac{\alpha_f C_f V_{dd}^2}{N_f \tau_g}. \tag{2.8}$$

$$P_{dynamic-l} = \frac{\alpha_l C_l V_{dd}^2}{N_l \tau_g}. \tag{2.9}$$

The equivalent gate delay is

$$\tau_g = \frac{kC_gV_{dd}}{(V_{gs} - V_{th})^a}, \tag{2.10}$$

where k and a are technology parameters. C_g is the total gate capacitance of a CMOS logic gate. For CMOS logic, $V_{gs} = V_{dd}$. For a latch-based design to attain the same frequency as flip-flop-based design, the scaled voltage (V_{ddl}) can be expressed in terms of the supply voltage of flip-flop-based design (V_{ddf}) as

$$N_l\tau_{g-Vddl} = N_f\tau_{g-Vddf}. \tag{2.11}$$

$$N_l\frac{V_{ddl}}{(V_{ddl} - V_{th})^a} = N_f\frac{V_{ddf}}{(V_{ddf} - V_{th})^a}. \tag{2.12}$$

It can be estimated from (2.12) how much voltage can be scaled for the latch-based design. The energy consumption ratio between the latch-based and flip-flop-based designs for the same operating frequency at different supply voltages in the super-V_{th} region is

$$\frac{E_l}{E_f} = \frac{\alpha_l C_l V_{ddl}^2}{\alpha_f C_f V_{ddf}^2}. \tag{2.13}$$

In the near/sub-V_{th} region, due to an exponential increase in delays the leakage energy consumption becomes significant. However, the strategy of scaling the supply voltage is not a good choice due to the high sensitivity of delay on supply voltage in the near/sub-V_{th} region. The frequency improvement in the latch-based design could significantly reduce the leakage energy consumption.

2.2 Latch-based design methodology

The method to convert a flip-flop-based design to a latch-based design is investigated in this section. The experiments that are performed in this section are based on an industrial 28-nm FDSOI CMOS technology. The standard cell libraries with regular threshold voltage (RVT) transistors are used.

2.2.1 Replace flip-flops by back-to-back connected latches

Latches typically consume lower power compared to flip-flops while displaying a speed advantage. This is also confirmed by the available data in the 28-nm FDSOI library that is used in this work. There are therefore chances to achieve power savings by

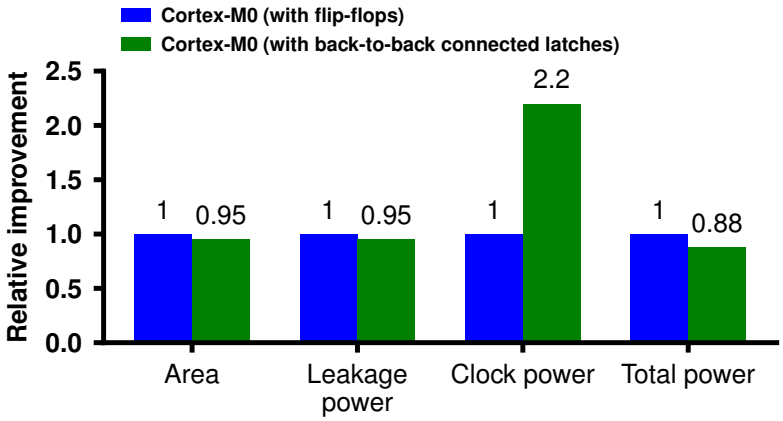


Figure 2.3. Comparison between the flip-flop-based design and back-to-back connected latch-based design for an ARM Cortex-M0 after backend physical design. All data are normalized to the data of the flip-flop-based Cortex-M0.

simply replacing flip-flops with back-to-back connected latches in a digital circuit. This transformation shows power savings of 7% for an ARM Cortex-M0 as shown in Fig. 2.3. Cadence RTL Compiler is used for logic synthesis, while Cadence Innovus Digital Implementation System is used for the backend physical design (placement and routing). A custom script is used to replace all the flip-flops in the design by back-to-back connected latches after the logic synthesis. In the back-to-back connected latch-based design, 18 clock buffers are required in the clock tree for driving twice the number of sequential elements, compared to 10 clock buffers for the flip-flop-based design. The clock tree power consumption is therefore increased from 14% of the total power consumption to 33% after the conversion to back-to-back connected latch-based design. The maximum frequency that can be achieved by the flip-flop-based design and back-to-back connected latch-based design is the same, as there is no time borrowing. By re-positioning the latches, there are chances to achieve higher performance by enabling time borrowing.

2.2.2 Flow for converting a flip-flop-based design to latch-based design

For converting a flip-flop-based design to a latch-based design, the flip-flops need to be split into master and slave latches and then retimed by using the commercial retiming

tools. The commercial EDA tools take the advantage of time borrowing property of latches and divide the combinational logic equally between the master and slave latches. Cadence RTL Compiler is used for this purpose in this work. Cadence RTL Compiler does not support the retiming of latch-based designs, but does support the retiming of flip-flop-based designs. Therefore, a work-around method is used to convert a flip-flop-based design to a latch-based design [105]. In the work-around strategy, the design is synthesized with a clock period T. Each flip-flop is replaced by two flip-flops. Then, the whole design is retimed at twice the synthesis frequency (clock period T/2). Since replacing a flip-flop by two flip-flops, the number of pipeline stages is doubled in the design. By balancing/splitting the combinational logic in the original pipeline stages of flip-flop-based design, the design with each flip-flop replaced by two flip-flops should be able to achieve twice the frequency. After retiming the circuit, the flip-flops are converted into negative and positive level-sensitive latches alternatively. After replacement with latches, the circuit is optimized for the required time period (T). Note that this process does not change the functionality of the circuit. This work delves deep in the process to convert any flip-flop-based design to a latch-based design, with power consumption as a minimizing trade-off factor for defining a suitable operating condition. The generic design flow for the proposed method is shown in Fig. 2.4.

2.2.3 New retiming strategy for converting a flip-flop-based design to latch-based design

While converting a flip-flop-based design to a latch-based design, there is a trade-off among the synthesis/retiming frequency, area, and timing slack, as illustrated in Fig. 2.5. When the frequency constraint is critical, the synthesis tool applies architecture change and over-sizing of gates to meet the timing (area increases) until it's impossible to meet the timing constraint. Sweeping the frequency from the point of timing slack 0 to the point when it's impossible to meet the timing is a large range. Therefore, to choose an optimum point for synthesis/retiming while converting a flip-flop-based design to latch-based design is an optimization problem. The optimization target in this work is for the maximum time borrowing. From (2.12) and (2.13), with more borrowed time, there could be wider supply voltage scaling for larger power savings. Note that for latch-based design, the activity factor is affected by the operating frequency as well due to glitches.

After replacing all the flip-flops by two flip-flops for a design synthesized at a relaxed frequency (slack $>>$ 0 ns) and retiming with clock constraint T/2, the combinational logic doesn't move properly as the retiming constraint is relaxed. Alternatively, when the synthesis frequency is relatively high, retiming results in relatively

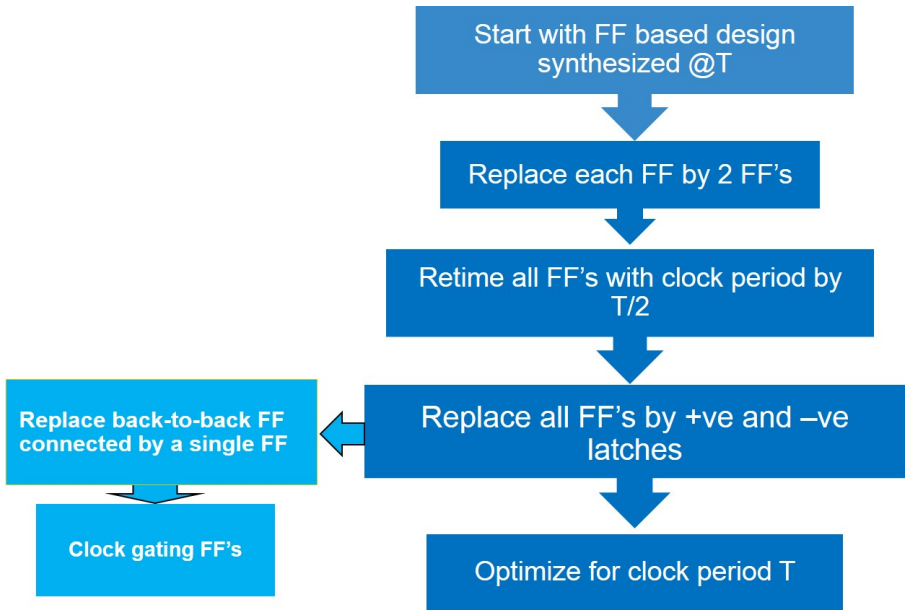


Figure 2.4. The flow for converting flip-flop-based design to latch-based design.

balanced pipeline stages. During this process of splitting logic between flip-flops, the synthesis tool adds additional flip-flops to maintain the functionality for branching of logic. The number of latches and gates after retiming and converting the initial flip-flop-based design with different synthesis frequencies to a latch-based design is shown in Table 2.1.

As listed in Table 2.1, when the circuit is synthesized/retimed at higher frequency, the number of latches added to divide the logic during retiming is large. The optimum operating condition for latch-based design which provides the maximum power savings could also be identified in Table 2.1. The circuit that is synthesized at 125 MHz has the maximum time borrowing capability, improving the performance by 41%, as compared to the flip-flop-based design. The results in Table 2.1 are based on the simulation after logic synthesis. The comparison of power consumption is to be done after physical design, which will be shown in Section IV.

From the logic synthesis results in Table 2.1, synthesizing and retiming the flip-flop-based design (FF design) to convert to latch-based design (LB design) at relatively high frequencies or relaxed frequencies lead to no performance improvement. There is an optimum frequency where the maximum performance enhancement or the largest power savings can be achieved for the latch-based design compared to the flip-flop-based design. As illustrated in Fig. 2.5, the optimum point is close to the point

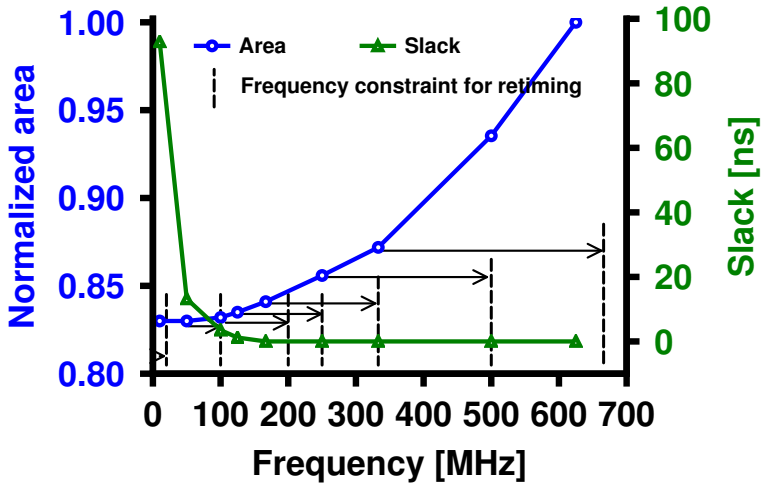
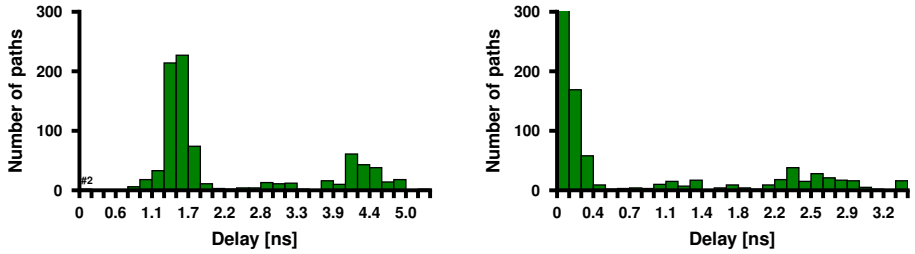


Figure 2.5. Illustration of the trade-off among the synthesis/retiming frequency, area, and timing slack for a latch-based Cortex-M0.

Table 2.1. Comparison of gate count, number of latches, and slack after retiming at different frequencies. Synthesis condition: Corner=Slow, V_{dd}=0.90 V, T=-40°C.

Cortex-M0	Synthesis (S)/ Retiming (R) Frequency	Gate Count	Flip-flop or latch number	Slack / Frequency
FF design	S=100 MHz (10 ns)	7147	841	3.5 ns / 153 MHz
LB design	R=200 MHz	7327	1732	4.4 ns / 178 MHz
FF design	S=125 MHz (8 ns)	7166	841	1.2 ns / 147 MHz
LB design	R=250 MHz	7422	1847	3.2 ns / 208 MHz
FF design	S=166.7 MHz (6 ns)	7284	841	0 ns / 166.7 MHz
LB design	R=333 MHz	7661	1986	1.1 ns / 204 MHz
FF design	S=250 MHz (4 ns)	7455	841	0 ns / 250 MHz
LB design	R=500 MHz	7819	2062	0 ns / 250 MHz
FF design	S=333 MHz (3 ns)	7594	841	0 ns / 333 MHz
LB design	R=666 MHz	8212	2220	0.2 ns / 357 MHz
FF design	S=500 MHz (2 ns)	8117	841	0 ns / 500 MHz
LB design	R=1 GHz	8634	2202	0 ns / 500 MHz

where the slope of area versus frequency plot is 1. By performing a few experiments, a small range that covers the optimum point can be identified. Afterwards, a sweep



(a) **(b)**

Figure 2.6. Path distribution of all the end points for Cortex-M0 synthesized at a) 166.6 MHz and b) 250 MHz.

of the frequency within this small range can be performed to capture the optimum frequency that provides the largest time borrowing, and hence the highest power savings compared to the flip-flop-based design. The latch-based design optimized for 125MHz and 166MHz shows the highest improvement in timing as compared to the flip-flop based design. To analyze the gain in the latch-based design, the distribution of path delays for all the end point of the Cortex-M0 synthesized using flip-flops at 166.6 MHz and 250 MHz are shown in Fig. 2.6. From Fig. 2.6a, it is clear that there is some margin for gaining time for the flip-flop-based design synthesized at 166 MHz. While for the flip-flop-based design synthesized at 250 MHz, the possibility of gains are limited.

While retiming the design where one flip-flop is replaced by two flip-flops, the division of logic depends on the number of gates between two stages. If there are limited logic gates in one stage, then the retiming does not work and eventually the latches remain back-to-back connected. We convert the back-to-back connected latches back to flip-flops. This results in a mixed design where latches are on the timing critical paths while flip-flops are on the non-critical paths. The number of latches and flip-flops in the mixed design (with Cortex-M0 as the test circuit) after the physical design is listed in Table 2.2. While converting the latch-based design synthesized at 100 MHz (achieved by retiming at 200 MHz) to a mixed design with flip-flops and latches, the mixed design has 819 flip-flops and 80 latches. This shows that retiming at relaxed frequencies doesn't divide the logic among the latches efficiently, and hence the design has very limited performance enhancement as compared to the design synthesized at 100 MHz. For the design synthesized at 250 MHz, the latch-based design does not result in performance improvements. However, the path is relaxed due to use of the latches, resulting in the decrease of number of combinational cells

Table 2.2. Comparison of gate count, number of flip-flops/latches, and slack after retiming at different frequencies for Cortex-M0 after backend physical design. Sign-off condition: Corner=Slow, V_{dd}=0.90 V, T=-40°C.

Cortex-M0	Synthesis (S)/ Retime (R) Frequency	Gate Count	Number of flip-flop/ latch	Slack (ns)/ Max Frequency(MHz)	Area (μm^2)	Leakage Power (nW)	Internal Power (μW)	Switching Power (μW)	Total Power (μW)	Clock Power (μW/%)
FF design	S=100 MHz (10 ns) R=200 MHz	7275	841 / -	3.4 ns / 151 MHz	8209	67	384	257	641	103 / 16%
LB design		7504	- / 1732	4 ns / 166 MHz	7903	69	243	374	618	241 / 39%
Mixed		6833	819 / 80	4 ns / 166 MHz	8355	74	401	276	677	115 / 17%
FF design	S=125 MHz (8 ns) R=250 MHz	7308	841 / -	1.3 ns / 149 MHz	8210	66	474	322	796	128 / 16%
LB design		7576	- / 1847	3.5 ns / 222 MHz	8020	70	315	482	798	319 / 40%
Mixed		7135	613 / 618	3.4 ns / 217 MHz	8425	76	462	392	855	194 / 23%
FF design	S=166.7 MHz (6 ns) R=333 MHz	7418	841 / -	0.3 ns / 175 MHz	8266	68	627	416	1044	172 / 16%
LB design		7777	- / 1986	1.4 ns / 217 MHz	8148	74	449	672	1122	477 / 42%
Mixed		7625	178 / 1578	1.4 ns / 217 MHz	8158	73	474	629	1104	405 / 36%
FF design	S=250 MHz (4 ns) R=500 MHz	7615	841 / -	0.2 ns / 263 MHz	8400	69	983	669	1653	267 / 16%
LB design		7944	- / 2062	0.1 ns / 256 MHz	8287	76	645	1020	1665	693 / 41%
Mixed		7909	193 / 1661	0.1 ns / 256 MHz	8457	77	769	960	1730	658 / 38%
FF design	S=333 MHz (3 ns) R=666 MHz	7740	841 / -	0 ns / 333 MHz	8465	72	1300	855	2155	350 / 16%
LB design		8308	- / 2220	0.1 ns / 345 MHz	8592	79	974	1459	2433	1067 / 44%
Mixed		8152	149 / 1880	0 ns / 333 MHz	8661	78	1057	1370	2427	969 / 40%
FF design	S=500 MHz (2 ns) R=1GHz	8303	841 / -	0 ns / 500 MHz	8831	82	1939	1353	3293	522 / 16%
LB design		8776	- / 2202	0 ns / 500 MHz	8812	84	1464	2274	3691	1536 / 42%
Mixed		8658	164 / 1854	0 ns / 500 MHz	8929	88	1576	2094	3671	1416 / 38%

Table 2.3. Comparison of power consumption by scaling voltage for Cortex-M0 after backend physical design. Corner=Slow, T=-40°C.

Cortex-M0	Slack/Max Frequency at 0.9 V	Voltage	Simulation Frequency	Leakage Power (nW)	Internal Power (μW)	Switching Power (μW)	Total Power (μW)
FF design	1.3 ns / 149 MHz	0.90 V	145 MHz	66	550	373	923
LB design	3.5 ns / 222 MHz	0.80 V	145 MHz	35	289	438	727
Mixed	3.4 ns / 217 MHz	0.80 V	145 MHz	37	420	352	772

used in the design. The decrease in the number of cells doesn't show any impact on power consumption, as it is over-shadowed by the increase in power consumption because of additional latches. For the design synthesized at 500 MHz and converted to latch-based design by retiming at 1 GHz, the latch-based design has 2202 latches which is 2.6× of the flip-flops in the flip-flop-based design. Alternatively, the mixed design that is synthesized at 500 MHz has 164 flip-flops and 1854 latches. This shows that the design is pushed for more duplicate paths and hence more latches are used due to the tight timing constraints. The latch-based design synthesized at 125 MHz and converted to a mixed design has a balanced result, showing 613 flip-flops and 618 latches. In this mixed design, indeed, the latches are on the critical paths and flip-flops are on the relaxed paths.

2.2.4 Evaluation of the proposed latch-based design methodology for super-V_{th} operation

The purely latch-based design and the mixed design with both latches and flip-flops are evaluated and compared with the flip-flop-based design in this section. The experimental results are based on the industrial 28-nm FDSOI CMOS technology. The worst-case corner is considered while evaluating the performance and power consumption of different designs. The ARM Cortex-M0 is used as the test circuit.

The comparison of the latch-based design, the mixed design, and the flip-flop-based design is shown in Table 2.2. Note that the switching power consumption in Table 2.2 is the power consumed by the interconnects and the primary ports of the standard cells, while the internal power consumption is the power consumed by the internal part of the standard cells. As listed in Table 2.2, the latch-based design converted from the flip-flop-based design synthesized at 125 MHz and retimed at 250 MHz has 48% improvement in frequency compared to the flip-flop-based design. The improvement in frequency can be used to scale the supply voltage for power savings. The supply voltage of the latch-based design is scaled to 0.80 V to achieve the same frequency (145 MHz) as the flip-flop-based design at 0.90 V. With supply voltage

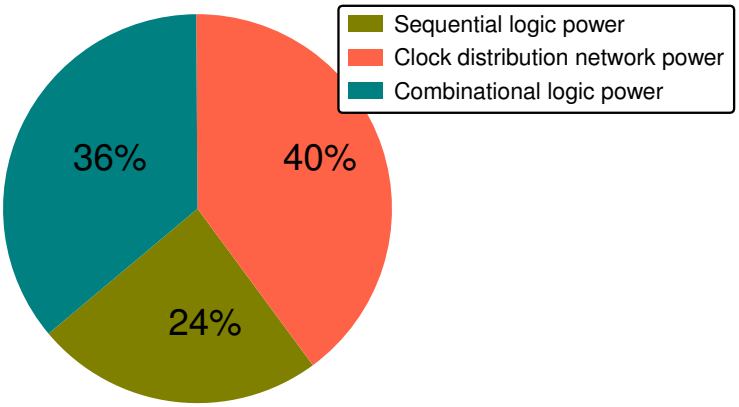


Figure 2.7. Power consumption breakdown for ARM-Cortex-M0 among sequential, combinational, and clock tree power consumption for latch-based design.

scaling, 21% power savings are achieved by the latch-based design as compared to the flip-flop-based design for the same performance. Furthermore, as listed in Table 2.3, the supply voltage scaling leads to 47% leakage power reduction with the latch-based design compared to the flip-flop-based design. It is interesting to note that the switching power consumption for the latch-based design is higher than the flip-flop-based design even after scaling the supply voltage. The latch-based design has higher switching power consumption because of more instances, nets, glitch propagation, and complex clock tree network as listed in Table 2.2. The number of clock tree buffers in the latch-based design clock tree is 21, whereas the flip-flop-based design has 10 clock tree buffers. Also, since latches are open for half of the clock cycle glitches produced in combinational logic of one stage can propagate to the next stage, if the design is pushed to operate in time borrowing mode.

In the latch-based design, it is observed that when the design is operated at 145 MHz and 0.80 V, because of time borrowing the glitches from one stage can propagate to the next stage. The glitch propagation from one stage to another stage results in more switching of the latches as well as the combinational logic. Flip-flops act as the filter of glitches [118]. To reduce the number of instances, nets, and glitches, the back-to-back connected latches are converted back to flip-flops. As listed in Table 2.2, the latch/flip-flop-mixed design synthesized at 125 MHz and retimed at 250 MHz has 45% improvement in frequency compared to the flip-flop-based design.

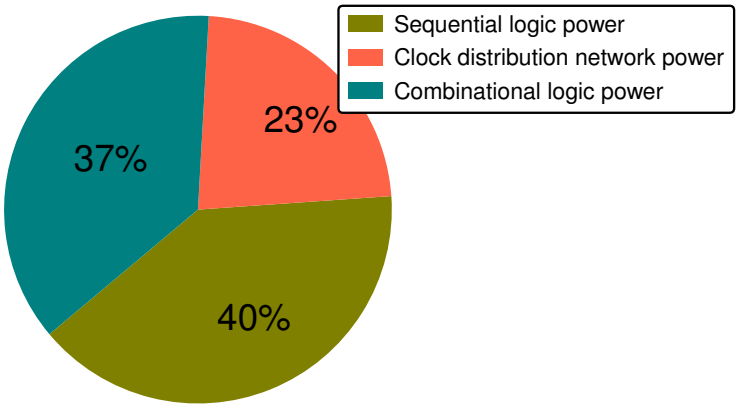


Figure 2.8. Power consumption breakdown for ARM-Cortex-M0 between sequential, combinational and clock tree power consumption for mixed design.

The latch/flip-flop-mixed design also achieves 16% power savings after scaling the supply voltage to 0.80 V as compared to the flip-flop-based design at 0.90 V, as listed in Table 2.3. Although the power savings for the mixed-design is lower as compared to the latch-based design, the mixed design serves as an important trade-off between the purely flip-flop-based design and the purely latch-based design. With the mixed design where flip-flops also exist, other low power techniques such as clock gating can be easily applied to the flip-flops. Alternatively, clock gating for latches is not trivial. Note that for fair comparison, no clock gating is applied for any of the designs that are evaluated in this work. Furthermore, whether clock gating is feasible or not and the effect of clock gating is heavily application dependent. Furthermore, the latch-based design synthesized at 166.7 MHz (retiming frequency is 333.3 MHz) shows 24% performance improvement, which results in supply voltage scaling to 0.85 V. The power savings for the latch-based design is 11% as compared to the flip-flop-based design at 0.90 V, as listed in Table 2.2.

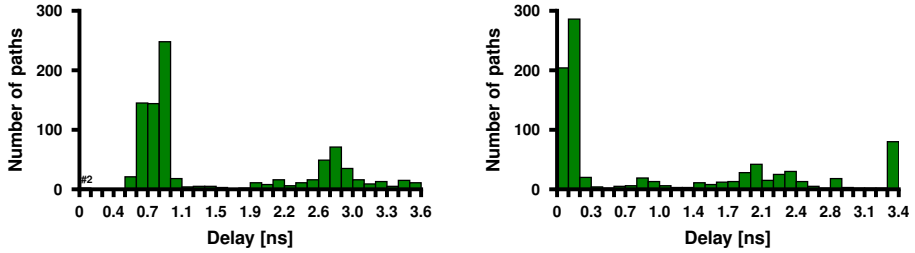
2.2.5 Evaluation of latch-based design at near/sub-V_{th} operation region

Voltage scaling to near/sub-V_{th} is required for saving energy for energy-constrained applications. In this section, the smart retiming for the latch-based design is eval-

Table 2.4. Comparison of near/sub-V_{th} performance improvement after retiming at different frequencies at slow corner, 0.45 V, and 0°C.

Cortex-M0	Synthesis (S)/ Retiming (R) frequency	Gate count	#FF or #Latch	Slack (frequency)	Leakage power (nW)	Clock power (μW)	Total power (μW)
FF design	S=5 MHz (200ns) / R=10 MHz	7147	841	30 ns (5.9 MHz)	67	1.3	13
LB design		8227	1822	41 ns (6.3 MHz)	69	3.6	18
FF design	S=6 MHz (166.7ns) / R=12 MHz	7615	841	0 ns (6.0 MHz)	69	1.4	13
LB design		8797	1962	53 ns (8.8 MHz)	75	6.0	19
FF design	S=14 MHz (71ns) / R=28 MHz	7791	841	0 ns (14 MHz)	84	3.2	31
LB design		9453	2204	4 ns (∼14 MHz)	86	10.6	35

uated for the near/sub-V_{th} region of operation. Unlike the latch-based design for super-V_{th} supply, it is not possible to scale voltage to save power in the near/sub-V_{th}, due to significantly large delay variation. The impact of global variations can be compensated by using techniques such as adaptive body-biasing and voltage scaling. These techniques are ineffective for local variations. In the state-of-the-art, two-phase latch-based pipelining is proposed to mitigate delay variability from local random variations [75]. A large time borrowing window of up to half a clock period can be achieved by a latch-based design which is advantageous in near/sub-V_{th} circuits with high process variations compared to flip-flop, soft-edge flip-flop and pulsed latch-based designs [119]–[121]. The proposed smart retiming strategy to convert a flip-flop-based design to a latch-based design is evaluated for the near/sub-V_{th} operation. The performance gains and power consumption using the smart retiming strategy for design synthesized at 5 MHz, 6 MHz, and 14 MHz (max) are shown in Table 2.4. In the near/sub-V_{th} region, the latch-based design's performance improvement using the smart retiming strategy follows the same trend as the super-V_{th} operation. The maximum performance gain is 46% for the design synthesized at 6 MHz. The significant performance improvement results in relatively higher leakage energy consumption savings. Although the leakage power increase by 9% for the latch-based design, the leakage energy saving is 21% for the latch-based design as compared to the flip-flop-based design. However, due to a significant increase in the clock-tree power consumption, the total energy consumption remains the same for the latch-based and flip-flop-based designs. At the maximum possible operating frequency for the flip-flop-based design, the smart retiming strategy result is negligible performance improvement with a 3× increase in the clock-tree power consumption overhead.



(a) **(b)**

Figure 2.9. Path distribution of all the end points after retiming for Cortex-M0 synthesized at a) 166.6MHz and b) 250MHz.

2.3 Discussions

2.3.1 Impact of retiming the flip-flop-based design

The methodology and the analysis to convert a flip-flop-based design to a latch-based design in this chapter is based on the using the retime option available in the Cadence Genus RTL compiler. The retime option of the tool works on balancing the paths between the flip-flops in the pipeline stages. The retiming of the flip-flop-based design synthesized at 166.6 MHz and 250 MHz is shown in Fig. 2.9. The retiming of flip-flop-based design improves the timing marginally, and it is not very effective. As shown in Fig. 2.9a when compared to Fig. 2.6a, although the distribution of path remains the same, the average delay of the maximum number of paths in the distribution decreases from 1.5 ns to 0.9 ns. The impact of converting a retimed flip-flop-based design to a latch-based design would result in a relatively lower performance improvement. Furthermore, the there is no timing impact of retiming on the design synthesized at 250 MHz. Additionally, the retiming step adds additional 49 and 40 flip-flops in the design, respectively.

2.3.2 Limitations of the proposed latch-based design methodology

The proposed approach of converting a flip-flop-based design to a latch-based design has some disadvantages. In terms of power savings, by trading the performance enhancement with supply voltage scaling, 21% power savings are achieved by the latch-based design as compared to the flip-flop-based design while operating in the super-V_{th} region. Note that the efficiency of the retiming strategy is the highest at an

optimum operating frequency point. If the operating frequency is too high or too low, then the advantage of the mixed latch-based design diminishes. For a flip-flop-based design operating at a maximum possible frequency (obtained after physical layout), the corresponding latch-based design doesn't achieve a higher operating frequency. The major drawback is that the achieved gains are within a limited range of frequency, which is relatively lower than compared to the maximum frequency achievable by the design. Additionally, the clock-tree design overhead results in a higher power consumption for the latch-based-design than that of the flip-flop-based design (see Table 2.2, 2.4). Furthermore, the choice between latch-based and mixed/latch-based designs is design-dependent. In latch-based designs, clock gating is not trivial. If the designer intends to take advantage of clock gating, the mixed design strategy is preferred. Additionally, the scan chain implementation for testing is a challenge for latch based designs. Usually, the latch-based designs are manually designed in pipeline stages to improve the performance. The EDA tools has limited support for timing and verification is case of latches.

2.4 Summary

In this chapter, an insight in terms of timing and power consumption for converting a flip-flop-based design to a latch-based design is revealed. A flow of converting a flip-flop-based design to a latch-based design, as well as a latch/flip-flop-mixed design is proposed. Based on a smart retiming strategy, the optimum operating condition for the latch-based as well as the mixed design is identified for achieving the maximum time borrowing and the highest power savings. The achieved gains are within a limited range of frequency which is relatively lower than compared to the maximum frequency achievable by the design. However, these gains are not feasible in the near/sub-V_{th} operation by the proposed automatic design methodology. The timing and power analysis are performed for a design operating in super-V_{th} as well as near/sub-V_{th} region.

Chapter 3

Design enablement for near/sub-threshold operation

The ongoing demand for reduction in energy consumption has motivated the design of digital circuits operating in near/sub-V_{th} region. Working in the near/sub-V_{th} region provides a very promising low power feature for applications with relatively low/medium performance requirements. The problem is that the performance is sensitive to process, voltage, and temperature variations while operating in the near/sub-V_{th} region. When designing digital circuits, these problems need to be considered. The standard library cells are the basic design elements made up of standard or macro functions that form the basis of digital design. Typically, the synthesis tools require enough freedom to improve the efficiency of the logic mapping and trade-off between power, delay, and area. Therefore, a wide variety of logic and sequential functions of different drive strengths are available in the standard cell library. A typical standard cell library consists of simple logic functions like INV, NAND, NOR, XOR, XNOR, MUX, etc., complex logic functions like half-adder, full-adder, AOI, OAI, etc., sequential cells like D-flip-flop, latches, scan-cells, etc., all with different drive strengths, and threshold voltage options. Additionally, the standard cell library also consists of some special cells like balanced rise and fall delay inverters and buffers for clock-tree implementations, delay elements for hold violations fixing, and low power cells like clock-gates, level-shifters, isolation cells, power-switches, state-retention flip-flops, etc.

The commercial standard cell libraries provided by the foundry are mostly designed and characterized for super-V_{th} voltage operation. Without any optimization, most cells do not have a robust operation in the presence of process variability at near/sub-V_{th} voltage. Therefore, for the near/sub-V_{th} operation an optimized standard cell library is required. In this chapter, an optimized standard cell library is developed for near/sub-V_{th} operating at a supply voltage of 0.4 V using the 28-nm

FDSOI CMOS technology. Naturally, cell optimization is technology-dependent and one has to take advantage of this fact. Therefore, an overview of the 28-nm FDSOI is also presented. Furthermore, we propose a systematic pruning methodology for the foundry-provided standard cell library. The proposed methodology determines the bad standard cells as per their relative degradation with voltage scaling.

3.1 Overview of 28-nm FDSOI CMOS technology

The 28-nm Ultra-Thin Body and Buried Oxide (UTBB) FDSOI CMOS technology is used in this work. The structure of FDSOI transistors is similar to traditional bulk devices [122]. The main difference is the thin insulator layer, also called the buried oxide (BOX), underneath the channel. Due to this, the channel thickness is also decreased compared to bulk silicon. The 28-nm FDSOI provides two variants: RVT and low threshold voltage (LVT) transistors. The RVT is based on conventional-well technology where the NMOS transistors reside within a P-well and the PMOS transistors reside within an N-well. While the LVT transistors are fabricated using a flip-well construction where the NMOS transistors are placed in an N-well and the PMOS transistors are placed in a P-well, respectively [123]. In LVT technology, the P-well is connected to the lowest voltage in the design (ground), which prevents the forward biasing of the body diode between the P-well and N-well. While it is possible to include both LVT and RVT cells in a single design, they need to be separated by a deep-N-well with a minimum spacing required in the layout. Additionally, the 28-nm FDSOI is less sensitive to process variations than conventional bulk CMOS and attractive for low voltage [124] as illustrated by designs in [125], [126], and [127]. FDSOI also provides the option of an ultra-wide range of forward and reverse body-biasing for LVT and RVT cells, respectively. The RVT reverse body-biasing (RBB) range goes from 0 V to -3 V and from +3 V to 0 V for NMOS and PMOS transistors, respectively. The default body-bias condition for RVT is the same as in the conventional well technology where the N-well is connected to the highest voltage and P-well is connected to the lowest voltage (GND) of the power domain. The LVT FBB range goes from 0 V to +3 V and from –3 V to 0 V for NMOS and PMOS transistors, respectively. In LVT, the default body-bias condition is when the N-well and P-well are connected to the lowest voltage (GND). In this work, LVT is selected for all cells, as the lower threshold voltage allows lower operating voltages. The cross-sectional view of the LVT transistors is shown in Fig. 3.1.

Another attractive feature of the 28-nm FDSOI CMOS technology is poly-biasing (PB). In PB, the minimum channel length (of 30 nm) can be extended by adding a PB mask layer, without changing the rest of the layout. In this technology, the PB mask layers are limited to 4 nm, 8 nm, 10 nm, and 16 nm. These PB mask-based

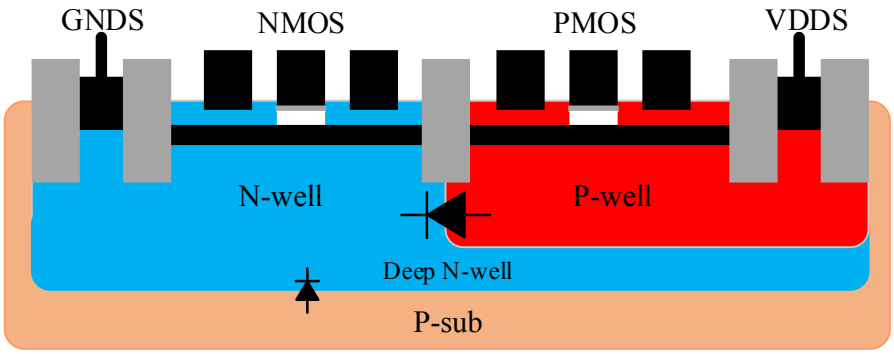


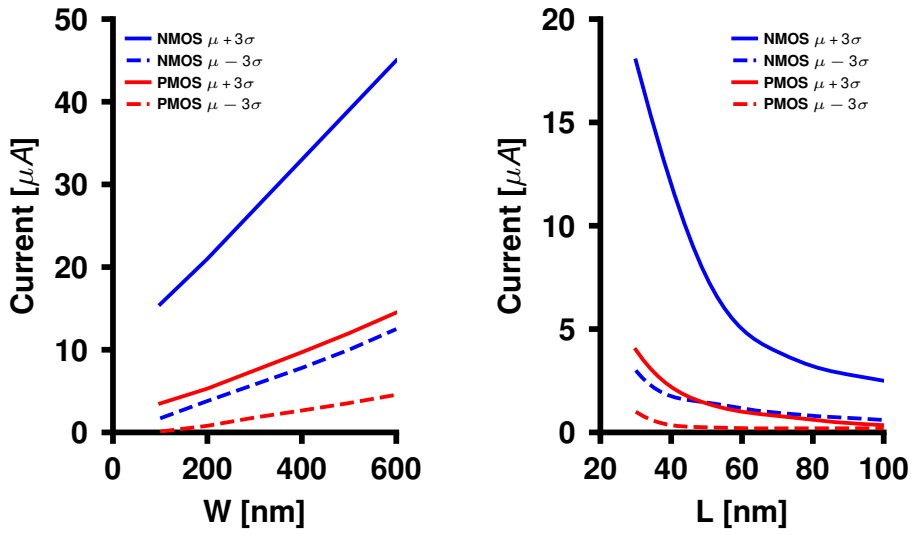
Figure 3.1. LVT transistors using flip-well technology cross-sectional view in 28-nm FDSOI

extensions do not increase the area of the transistors. For other channel lengths, the poly length has to be adjusted along with the source and drain area in the layout, but this hinders the regularity of gate spacing found in the standard cells and increases the area. Additionally, the transistors connected by their active regions in the layout are required to use the same PB. Increasing the channel length can be used to reduce leakage currents at the cost of performance. For this purpose, the existing commercial library provides variants of the standard cells for all PB values. However, these variants simply apply the PB to the entire cell without adjusting the sizing to account for the change in W/L ratios.

In the following section, a brief review of the standard cell design techniques for near/sub-V_{th} operation is presented. Additionally, the current variations because of length tuning, width tuning, and the impact of process variations are also studied.

3.2 Standard cell library sizing literature review

The logic gates exhibit DC failures or show extreme delay degradation due to reduced transistor on/off current ratios and increased sensitivity to process variations [27], [33], [128], [129]. Therefore, a careful design of standard cells working in near/sub-V_{th} has been pursued. For instance, in [130], the fundamentals of logical effort in the near/sub-V_{th} region were developed for the sake of optimal device sizing. In [44], [52], [77], [129], [131]–[133], different logic design and sizing techniques are proposed. In [44], [129], [132], [133], a variation aware sizing methodology is used. Alternatively, in



(a) Width tuning (L=30nm). **(b)** Length tuning (W=160nm).

Figure 3.2. Effect of width and length tuning on PMOS/NMOS on-current for $|V_{GS}| = |V_{DS}| = 0.40\,\text{V}$, T=27°C. The plots show the lower and upper current bounds.

[52], [77], a transmission gate logic design strategy is used. Over the years, different procedures to optimize the cells for near/sub-V_{th} have been developed. In [134], a library for 28-nm FDSOI has been developed for sub-V_{th} supply voltage range by using body biasing as the main design variable and tested on a MAC block, showing a 21% reduction in energy consumption compared to other works.

The 28-nm FDSOI standard cell library from the foundry is designed and optimized for a nominal operating supply voltage of 0.9 V. The performance of the standard cells is not optimal in the near/sub-V_{th} region. Firstly, the on-current behavior of the NMOS and PMOS transistors is characterized by process variations. The two design parameters that were varied are the width and length of the transistors. The effects of width and length tuning for the NMOS and PMOS transistors operating at 0.4 V and a temperature of 27°C are studied to gain a better understanding of the technology.

Impact of width tuning: To understand the impact of width tuning of an LVT NMOS transistor, the gate, and the drain are connected to VDD of 0.4 V, whereas

both source and body are connected to the ground. For an LVT PMOS transistor, the source is connected to VDD of 0.4 V, whereas gate, drain, and body are connected to the ground. The width tuning impacts the threshold voltage and drains current. For the transistors, the length is kept constant at the minimum of 30 nm, and the width is varied in the step of 10 nm from the minimum width of 80 nm up to 600 nm. As mentioned earlier, at near/sub-V_{th} operation, the process variations have a major impact on the performance of the transistor. For each set of transistor parameters, 250 Monte Carlo simulations are run. The simulated drain currents are plotted in Fig. 3.2a. The plotted lower and upper bounds of the distributions are defined as μ - 3σ and μ + 3σ, respectively. From the σ and lower/upper bounds plots, it is observed that increasing the width decreases the variation. This is expected, as process variations are relatively smaller with a bigger device. The current plots show that both the mean and the variation increase with increasing width, as the current is dependent on the W/L ratio. Additionally, the size of PMOS needs to be about $5\times$ (W_p/W_n) that of an NMOS to balance the on-currents. This asymmetry is problematic for layout purposes. Therefore, it is imperative to include length tuning in the balancing strategy.

Impact of length tuning: To understand the impact of length tuning of an LVT NMOS transistor, the test setup is kept the same as for the width tuning. However, the width is chosen to be a constant of 160 nm, whereas the length is swept from 30 nm to 100 nm in steps of 5 nm. Similar to width tuning, a 250 point Monte Carlo simulation was executed to plot the drain current variation with length tuning. Fig. 3.2b shows that the current decreases monotonically with increasing length, as V_{th} increases and the ratio W/L decreases. For 28-nm FDSOI, the options for length tuning are more limited, as increasing width increases current and its variation, whereas increasing length decreases both. The length tuning options are only by 4 nm (PB4), 8 nm (PB8), 10 nm (PB10), and 16 nm (PB16) without increasing the area of the cell and without compromising the poly pitch symmetry between PMOS and NMOS transistors.

3.3 Sizing Methodology

The 28-nm FDSOI CMOS technology provides control over the threshold voltage of the transistor by poly-biasing. From the experiments in the previous section, the following simple balancing strategy was devised. Essentially to balance the PMOS and NMOS poly-biasing is used either on the NMOS or the PMOS transistor. The approach balanced the worst rise transition time against the worst fall transition time to reduce delay spread. The slack available in the best-case timing arc is reduced by using poly-biased transistors on that path, while the timing of the worst-case timing

arc is improved by using up-sized transistors. The length of all PMOS transistors in the standard cells was kept constant at the minimum of 30 nm, as there is a massive difference in the on-currents of NMOS and PMOS. The NMOS gate length is varied by only using PB to keep the regularity of the standard cells intact. The gate length is in the range of 30 nm (PB0), 34 nm (PB4), 38 nm (PB8), 40 nm (PB10), and 46 nm (PB16). The range of transistor widths was determined from the existing commercial libraries. The foundry offers two variants 8-track and 12-track library. The tracks indicate the height of the cells which shows how many horizontal metal tracks can fit in the height of the cell. Each track corresponds to the minimum pitch of M1 (100nm). The 8-track library is advertised as a low-power variant. Our goal is also to design a low-power library, hence the 8-track library is considered as a reference. In the 8-track library, a single NMOS transistor can have a maximum width of 222 nm and a single PMOS can be 322 nm while following all the design rule constraints. Therefore, the range of NMOS widths was chosen to be from 80 nm to 220 nm. For PMOS widths, it was chosen to be from 160 nm to 320 nm. This decision is made to keep the design space within the limit and to save computation time. The newly designed cells have the same height with the same N-well and P-well extension, making them symmetric and compatible with the foundry-provided standard cell library. The newly designed cells have the same poly pitch of 136 nm the same as in the foundry provided standard cell library. The foundry provided standard cells are not design rule chek (DRC) clean standalone due to the poly extension on the top and bottom side. The constant poly pitch of 136 nm is important to make the final cells DRC clean in an SoC, when they are placed in alternate tracks. The main contribution of this work is balancing the worst rise and fall delays without changing the height of the cells, poly separation, area, and still compatible with the existing standard cell library.

The sizing methodology focuses on balancing the pull-up network (PUN) and pull-down network (PDN) to improve the robustness of the cells by adjusting the sizing. The worst rise transition is balanced against the worst fall transition for each cell to reduce the delay spread. As a result, the slack available in the best-case timing arc is reduced by using poly-biasing of transistors on that path, while the timing of the worst-case timing arc is improved by using bigger transistors. In this method, although the worst case is improved, the best case transition delays are also increased. This results in the difference between all transition delays shrinking, making the cell more balanced. A simple testbench is used for optimization. At each input of the design under test (DUT), a voltage source was connected with an input transition time of 10 ns. A load capacitor of 10 fF was connected at the output of DUT. These values were estimated from the existing characterized library. All body connections were connected to the ground, as all transistors used were LVT. The supply voltage for the design optimization is 0.4 V. The sizing optimization was performed at the

Transistor	Existing cell sizing	Balanced cell sizing
P1	164 nm/30 nm	320 nm/30 nm
P2	164 nm/30 nm	320 nm/30 nm
N1	108 nm/30 nm	80 nm/46 nm
N2	108 nm/30 nm	80 nm/46 nm

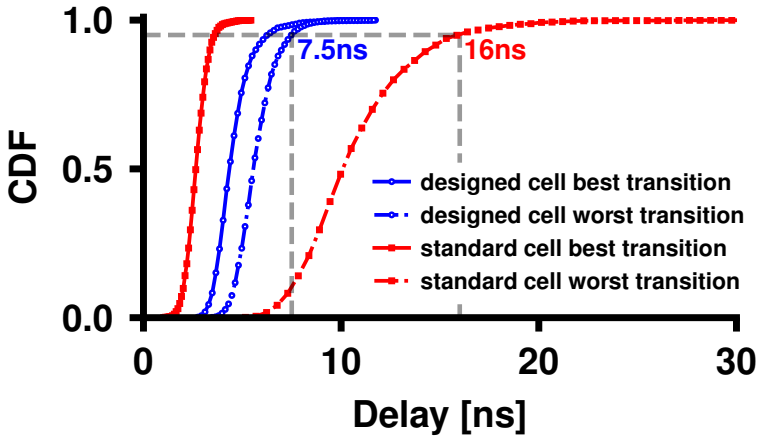


Figure 3.3. Transistor sizing and CDFs of best and worst transition delays using the new and the PB0 libraries for NOR2 at VDD = 0.4V, T = 25° C.

typical (TT) process corner. The results of sizing optimization were verified by Monte Carlo runs for every test to verify the yield improvement. The reference is chosen to be the foundry delivered 8-track LVT zero poly-bias (PB0) library. The optimization is done such that the overall cell area remains constant with regard to the area before optimization.

3.3.1 Combinational cells

The designed new library consists of 22 optimized combinational cells. To illustrate the effect of the balancing, for a minimal drive strength NOR2 cell the cumulative distribution functions (CDFs) of the best and worst transition delays of 250 Monte Carlo simulations are shown in Fig. 3.3. The transistors are listed from top to bottom where P1 is the top transistor in a conventional schematic. At the 90% yield point, the gap reduced from 16 ns to only 7.5 ns. To achieve this, the PMOS transistors are up-sized and PB16 is used for the downsized NMOS transistors.

Transistor	Existing cell sizing	Balanced cell sizing
P1	164 nm/30 nm	300 nm/30 nm
P2	164 nm/30 nm	300 nm/30 nm
N1	108 nm/30 nm	125 nm/30 nm
N2	108 nm/30 nm	170 nm/30 nm

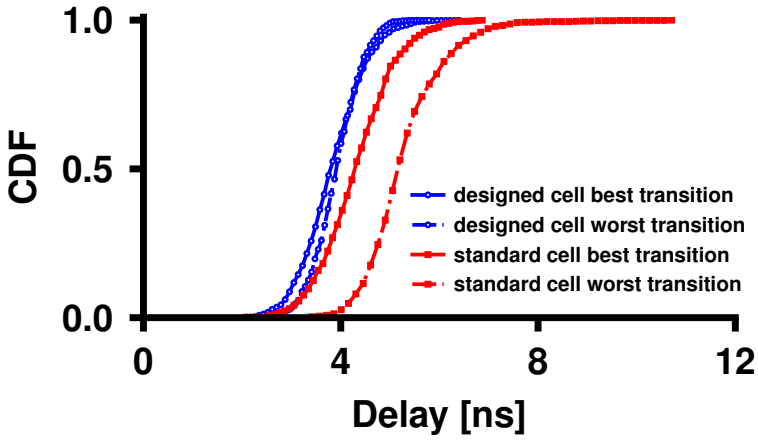


Figure 3.4. Transistor sizing and CDFs of best and worst transition delays using the new and the PB0 libraries for NAND2 at VDD = 0.4V, T = 25° C.

For a minimal drive strength NAND2 cell the CDFs are shown in Fig. 3.4. In the NAND2 cell, N2 is the bottom transistor, and N1 is the top one. To balance the cell, both the NMOS and PMOS are up-sized, and the lengths of both NMOS are kept minimal. The up-sizing of the PMOS improves the original worst transition delay (rise transitions). However, to balance this rise transition with the worst fall, the NMOS transistors are up-sized. As compared to other cells, the worst rise paths contain an equal number of PMOS in series as NMOS in series or more. Therefore, the NMOS transistors are usually downsized for the other cells, such as the NOR2 and inverter. As shown in Fig. 3.4, certainly the best and worst transition delays are much closer as compared to the standard cell library. The mean of worst delay is ∼5.5 ns, whereas the mean of best rise time is ∼4.2 ns for the NAND2 cell from the foundry. Therefore, the difference between the mean of worst rise and fall delays is ∼31%. The difference between the mean of worst rise and fall transitions is negligible. At the 90% yield point, the difference between the worst and best transitions for the balanced NAND2 cell is also negligible, compared to ∼2ns for the existing PB0 library.

Table 3.1. D flip-flop sizing results.

	Transition	**Existing cell**	**Balanced cell**
CLK-Q delay	0 –> 1	839 ps	767 ps
Setup time	0 –> 1	105 ps	232 ps
CLK-Q delay	1 –> 0	983 ps	732 ps
Setup time	1 –> 0	354 ps	315 ps
Worst CLK-Q delay + setup time		1337 ps	1082 ps

3.3.2 Sequential cells

The D flip-flops (DFF) are the basic requirement for synthesizing any digital synchronous design. For a DFF, the optimization is different from that of combinational circuits. For the DFF, the relevant timing characteristics are clock-to-Q (CLK-Q) delay, setup time, and hold time. The performance of a DFF is determined by the sum of the CLK-Q delay and the setup time. Therefore, the sum of the worst CLK-Q delay and worst setup time is optimized. There are many transistors in a DFF, therefore the search space for optimum sizing of transistors is huge. To reduce the search space some recommendations from [135] were used. The recommendations are not to change the feedback transistors, as they are almost sized minimally and barely impact the delay and the setup time. The transistors in series are sized the same. The setup time is defined the same as in [136], the data to clock offset where the CLK-Q delay is increased by 5% of its nominal value. The nominal CLK-Q delay was determined by measuring the CLK-Q delay when the data to clock offset is very large, as the CLK-Q delay is stable for sufficiently large offsets. In this work, 10 ns was determined to be enough. The data to clock offset was then increased until the 5% increase point is reached. This procedure is also described and shown in [137]. Most of the widths of the NMOS transistors are either untouched or smaller for the optimized cell, compared to the library cell. Table 3.1 shows the difference between the existing library DFF and the optimized DFF operating in the typical corner at 27°C. In the library cell, the falling CLK-Q delay is 17% higher as compared to the rising CLK-Q delay showing a big difference between the rise and fall transitions. The setup times differ by a factor of 3.4×. For the optimized DFF, there is only a 5% difference between the CLK-Q delays. The setup times only differ by a factor of 1.4. The worst-case sum of delay and setup time is also 20% less. This can be attributed to the smaller transistors, in general, reducing the time required to charge internal nodes.

The layout of the optimized standard logic cells and sequential cells are generated automatically. For automatic layout generation, the schematics are represented

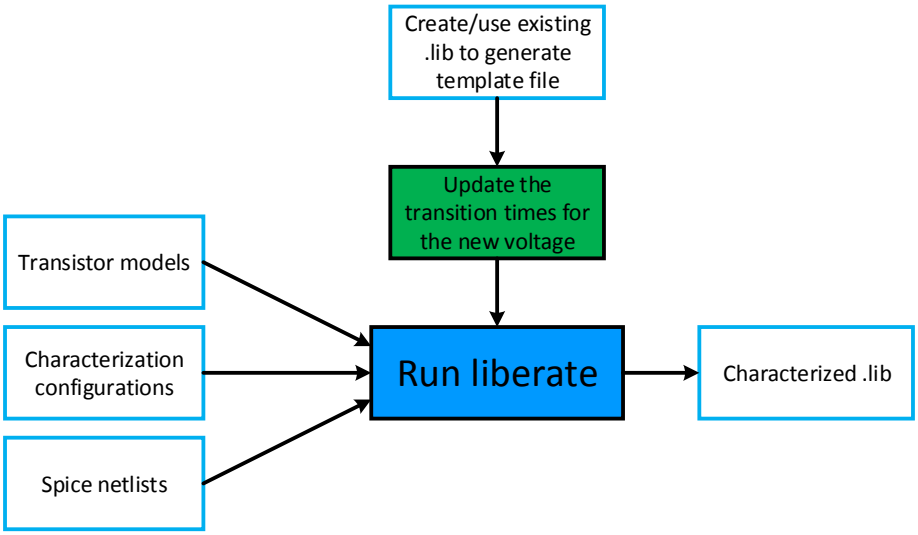


Figure 3.5. Standard-cell library characterization using Cadence Liberate.

as graphs and the concept of the Euler path is used to determine the optimal gate ordering. Afterward, maze routing algorithms are used to route the nets. The generated layout is extracted using Cadence Quantus QRC and characterized to generate liberty files for different corners and temperatures.

3.4 Near/sub-V_{th} standard-cell library characterization

The characterization of the library is an important step to enable the usage of the cells for HDL synthesis and place & route tools. Characterization of the library aims to generate timing and power models for each cell. Designing especially at advanced nodes requires multiple library PVT views to avoid failure due to insufficient/inaccurate sign-off. For accurate modeling of voltage variation or temperature gradients, it is important to characterize each library at multiple corners, multiple voltages, and multiple temperatures. The existing characterization tools have automated most of the library generation process. These tools run simulations under realistic conditions for all possible timing, power arcs. The simulation results are extracted and compiled in the libraries. In this work, Cadence Liberate is used for

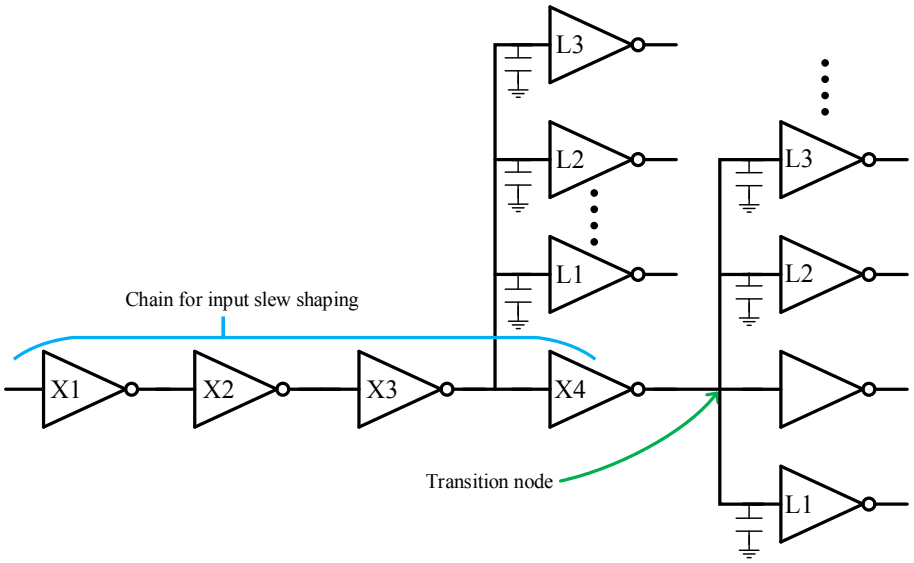


Figure 3.6. Testbench for calculating the transition slew for different fan-out ranges.

the standard cell characterization. The cells are characterized in the most popular non-linear delay model (NLDM). NLDM uses a voltage source with the appropriate impedance and a simple load model. The characterization for the near/sub-V_{th} operation requires an additional effort in template creation.

The standard characterization flow requirements and overview are shown in Fig. 3.5. It requires the spice netlist extracted from the cell layout, transistor models provided by the foundry, template file, and characterization configuration file. The output of the characterization is a standard liberty file. The template file required for characterization contains the definitions of the cell, timing, and power templates that will be used for characterization. The template file contains all input to output timing arc definition and for each timing arc, it contains power definitions. The standard cell delay and power consumption are a function of the input signal transition time (both rise and fall) and output load capacitance. For the standard cell libraries, the template file can be generated using the existing liberty files. The near/sub-V_{th} voltage characterization requires modification of the transition (slew) in the template file according to the supply voltage. The maximum slew is calculated using the sample design in Fig. 3.6. A sample design is created using the inverters or buffers of minimal driving strength to generate the transition timetable. The fan-out load capacitances

are modeled for the interconnects. The typical maximum fan-out of a standard cell is from 50 to 100 when they operate at super-V_{th} region. However, at near/sub-V_{th} buffering is preferred to drive large fan-out loads. The input fan-out of the DUT is considered half of the output fan-out. The slews are simulated by varying the number of cells from 1 to a fan-out of 100 at the output node of the DUT. In the generated template file, the slew values are replaced with the newly generated slew values. Then this new template file is used for the characterization of the library at near/sub-V_{th} voltage. During synthesis, the fan-out limit can be set to be in the middle of the characterized delay table. At super threshold supply voltage, the fan-out can be as high as 100 as the drive strength of the cells are high. For near/sub-V_{th} operation low fan-out is preferred as the cell drive strength is significantly lower.

3.4.1 Comparison with foundry PB0 library

Finally, the comparison of the designed library with the chosen foundry delivered 8-track LVT zero poly-biasing (PB0) library is shown in Fig. 3.7. Most of the designed cells are below the unity line. This shows that the worst-case delay of the designed cells is less for the new library compared to the PB0 library. Since the sizing of the cells is aimed to balance the PUN and PDN thereby improving the worst-case transition at the cost of the best case transition. The cells on or above the unity line all have maximum width PMOS transistors in the PB0 library. The worst-case transition can therefore not be improved without violating the constraints. The minimum sized NOR2 gate had the most room for improvement, see Fig. 3.7.

3.5 Experimental results

For evaluating the design library with the existing library. The synthesis of multiple benchmarks is performed. For synthesis, the worst conditions are used with an operating supply of 0.36 V (10% lower than nominal), SS corner, and temperature of 0°C. The PB0 and PB4 commercial libraries were also re-characterized at these conditions.

As a benchmark, first, an ARM Cortex-M0 is synthesized over a clock period ranging from 70 ns to 200 ns using Cadence Genus. At first, the resulting synthesized circuits are compared with the synthesis result of the commercial library in terms of minimum clock period and leakage power consumption savings. For a fair comparison, only similar cells with the same area as the designed cells are used for synthesis from the commercial library. For the commercial libraries, three different sets of libraries are used. 1) Only PB0 library, which is expected to achieve maximum performance. 2) A mixture of both the PB0 and the PB4 library. The PB4

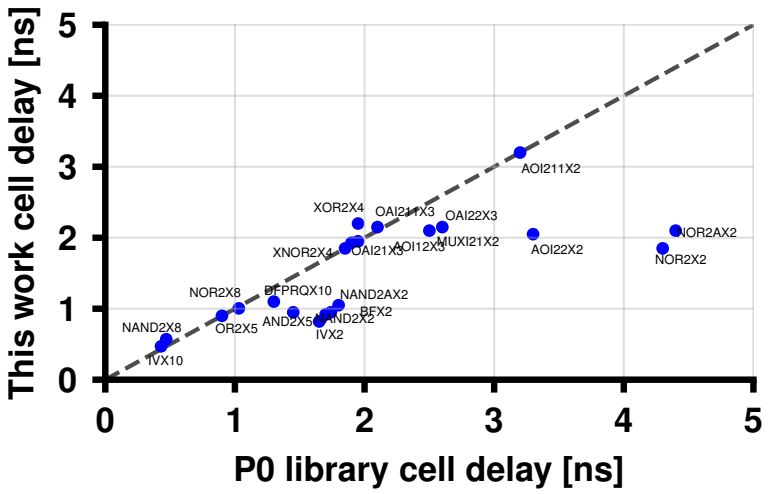


Figure 3.7. Worst case propagation delay comparison between the PB0 and the balanced libraries at V_{dd}=0.4 V, TT corner, and 25°C.

library cells are expected to be used on the non-critical paths to reduce leakage power consumption. 3) Only PB4 library, which is expected to have lower leakage power consumption as compared to the PB0 library. The comparison of leakage power consumption as a function of clock period for synthesized ARM CM0 is shown in Fig. 3.8. The detailed comparison is tabulated in Table 3.2.

Overall, the maximum performance achieved is slightly degraded for the proposed library as compared to the PB0 library. The PB0 with/without the PB4 library could achieve a maximum clock period of 75 ns, whereas the proposed library could only achieve 80 ns (higher by about 7%). However, the achieved clock period is about 39% lower as compared to the low leakage PB4 library, which could achieve 130 ns.

Comparing the leakage power consumption, the proposed library can achieve 35-40% lower leakage power as compared to the PB0 library, see Fig. 3.8. Essentially, the proposed library consists of smaller width and larger length NMOS transistors than the PB0 library, therefore, this result is expected. The mixture of the PB4 and PB0 libraries consumes lower leakage power as expected. Therefore, between 80 ns and 95 ns of the clock period, leakage power saving of up to 20% is achieved. The leakage power consumption drops for the lower target clock period as the fraction of cells selected from the PB4 library increases gradually. Compared to only the PB4 library, The leakage power consumption for the proposed library is 50-70% higher.

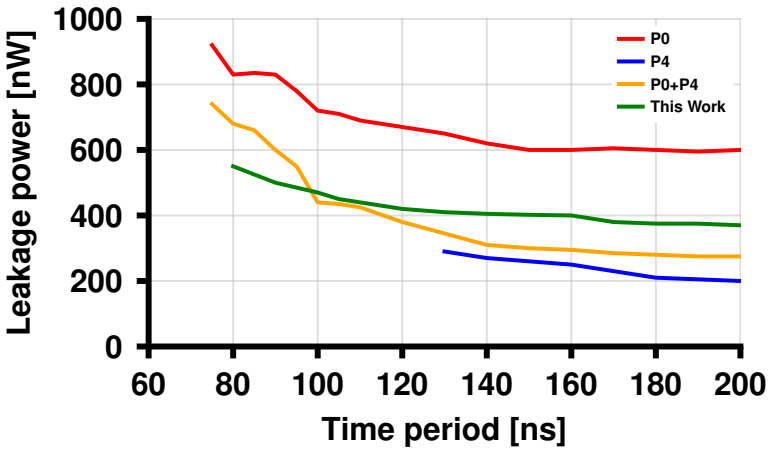


Figure 3.8. Leakage power and achieved frequency of the synthesized ARM Cortex-M0 using different libraries.

The total area of the synthesized circuit of Cortex-M0 is the same as when synthesized using the PB0 library. The mixture of PB0 and PB4 libraries consumes an area up to 5% higher than the designed library, due to a higher number of cells used. The use of only the PB4 library results in 10-20% more area, as more cells are required for buffering.

The dynamic power consumption is 5-10% lower in the clock period range between 80 ns and 100 ns as compared to the PB0 library. For the majority of the cells, the PMOS widths are increased, whereas the NMOS widths are decreased and lengths are increased. Non-critical transistors in the complex logic gates are also downsized. Overall, the capacitances are smaller, resulting in a reduction of dynamic power consumption. Also, in IoT applications, where the device is not always active, leakage power plays a bigger role in the overall energy consumption.

For benchmarking, the ARM Cortex-M0 and three different ITC benchmarks are synthesized using the proposed library, as well as combinations of the PB0 and PB4 commercial libraries. The synthesized circuits are compared in terms of area, minimum period, leakage power, and dynamic power. The results of the synthesis are shown in Table 3.2.

The minimum achievable clock period using the proposed library is 7% lower than that of the PB0 library for the Cortex-M0. However, at 80 ns, the leakage power consumption using the designed library reduces by 38% and the dynamic

Table 3.2. Comparison of the designed new standard cell library with the existing library for ARM Cortex-M0 (CM0) and ITC benchmarks post synthesis using Cadence Genus-RTL compiler for VDD = 0.36 V, SS corner, T = 0°C.

Design	Libraries	Minimum period [ns]	Comparison period [ns]	Area [μm^2]	Leakage power consumption [nW]	Dynamic power consumption [μW]
CM0	PB0	75	80	12011	885	35.7
CM0	PB0+PB4	75	80	12821 (+7%)	750 (-15%)	37.0 (+4%)
CM0	New	80 (+7%)	80	11840 (-1%)	547 (-38%)	32.7 (-8%)
CM0	New+PB0+PB4	75	80	11735 (-2%)	561 (-37%)	32.6 (-9%)
b18	PB0	65	75	40173	2422	99.1
b18	PB0+PB4	65	75	41403 (+3%)	1447 (-40%)	97.3 (-2%)
b18	New	75 (+15%)	75	41012 (+2%)	1546 (-36%)	97.8 (-1%)
b18	New+PB0+PB4	65	75	40543 (+1%)	1212 (-50%)	92.2 (-7%)
b20	PB0	65	75	8544	628	39.1
b20	PB0+PB4	70 (+8%)	75	9786 (+15%)	554 (-12%)	42.7 (+9%)
b20	New	75	75	9447 (+11%)	439 (-30%)	39.9 (+2%)
b20	New+PB0+PB4	65	75	8613 (+1%)	419 (-33%)	37.1 (-5%)
b22	PB0	65	75	12603	918	52.5
b22	PB0+PB4	70 (+8%)	75	14549 (+15%)	824 (-10%)	57.9 (+10%)
b22	New	75	75	14103 (+12%)	653 (-29%)	53.7 (+2%)
b22	New+PB0+PB4	65	75	12875 (+2%)	623 (-32%)	49.3 (-6%)

power consumption reduces by 8% for the almost same area, compared to the PB0 library. The PB0 and PB4 combination decreases the leakage power consumption by 15%, but increases the dynamic power consumption by 4% and area by 7% as compared to the PB0 library. Adding the proposed library to PB0 and PB4 results in similar power consumption reduction as with only the new library without any performance penalty. The synthesis using the combination of the proposed library with PB0 and PB4 library shows that the proposed library is more preferred by the tool than other libraries as shown in Fig. 3.9.

For the ITC benchmarks, the new library shows 29-36% reductions in leakage power for a 15% decrease in maximum performance. For b18, PB0+PB4 actually provides an additional 4% leakage power reduction compared to the new library. However, combining all three libraries results in the least leakage power without any decrease in performance. The dynamic power is also reduced by 7%. Similar reductions are observed for the other ITC benchmarks. The area is minimal for the PB0 library, although the combination of all three libraries only increases the area by 2% at most.

To demonstrate the improvement in robustness of the designed library, the critical

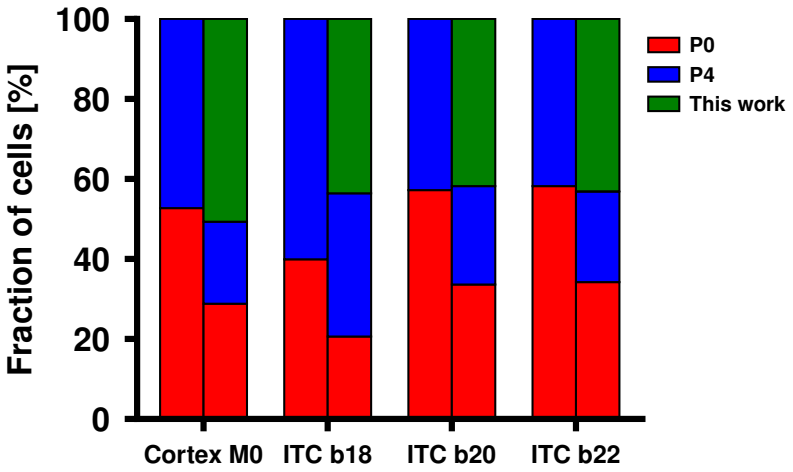


Figure 3.9. Library cell distribution for synthesis using combination of the proposed library with PB0 and PB4 commercial library.

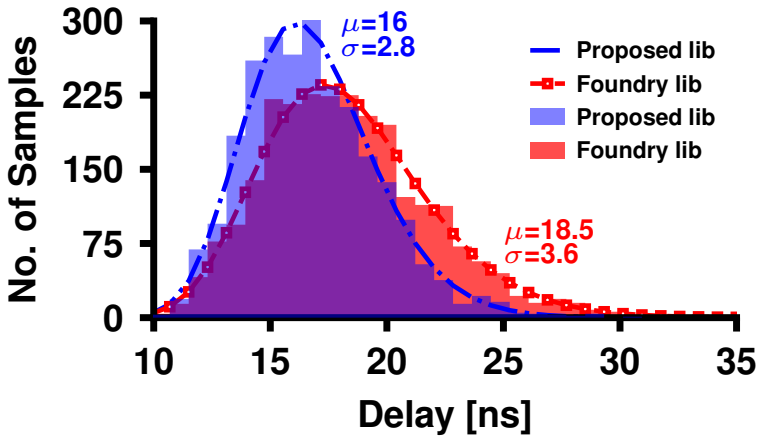


Figure 3.10. Normalized histograms of the critical path delay for ARM Cortex-M0 synthesized for 80 ns at 0°C, VDD = 0.4 V.

path is extracted from the synthesized ARM Cortex-M0 for a target period of 80 ns. From the commercial libraries, the critical path consists of only PB0 standard cells. 2500 Monte Carlo simulations (both global variations and local mismatch are enabled) are run to determine the delay spread. In Figure 3.10, the normalized histograms of the critical path delays are presented for VDD = 0.4 V. For the PB0 library, μ is 18.3 ns, whereas, for the new library, μ is only 16.8 ns, about 8% lower. For σ, it is 3.6 ns versus 2.8 ns, a 22% decrease. Thus, the variation (σ/μ) decreases from 19.6% to 16.7%. At the 90% yield point, the delay decreases from 22.9 ns to 20.3 ns, an 11% reduction. At the 99% yield point, the difference is even bigger, from 28.1 ns to 24.2 ns, a 14% decrease. The designed library performs relatively better as compared to the foundry-provided library in terms of leakage power consumption and process variations operating in the near/sub-V_{th} region.

3.6 Standard cell library pruning

Usually, a significant effort is required to design a new standard-cell library. One has to acknowledge that often enough the commercial library meets the functional yield constrain up to a certain voltage limit in sub-V_{th}. However, some cells have a large driving-current variability, which can remarkably deteriorate the timing. Often it is possible to prune cells from the library which show relatively large variability in speed and drive current when the voltage is scaled to near/sub-V_{th}. In [27], [34], [47], [68], generic guidelines for pruning of the commercial library cells are explored. Those guidelines suggest to *1)* avoid more than 3-stacked transistors such as 4-input cells, because they introduce large current variability, *2)* to avoid ratioed cells since the correct functioning of ratioed cells largely depends on correct sizing of transistors, even a small variation in V_{th} could show large current variability on the active or leakage current, *3)* to avoid cells with transistor sizing dependent functionality, and *4)* to avoid logic cells with feedback, the feedback is usually positive and the operation depends on the loop gain which changes with V_{th} variations, the output could have stuck at 0/1 failure. However, these guidelines are ad hoc and do not provide a good insight into performance/degradation differences among cells. Furthermore, in 28-nm FDSOI there are no cells with greater than 3-stacked transistors. Some 4-input cells are designed using three 2-input gates. Given the difference between the PMOS and NMOS, the guidelines in the literature recommend removing cells consisting of greater than 2-PMOS transistors in series.

We show in this work a new and more involved and reliable methodology for pruning library cells, which significantly degrade at sub-V_{th} voltages. The proposed methodology is based on the rate of delay change with voltage scaling. For illustration, we use the RVT library from the 28-nm FDSOI technology, already character-

Algorithm 3.1 Algorithm for library pruning

procedure PRUNELIB($0p9v.lib,\ 0p4v.lib$)
  Read the characterized super-V_{th} and sub-V_{th} Liberty files
  for each cell $i \in CellList$ **do**
    $\tau_{super-V_{th}}(i)$= worst delay of all timing arcs from 0p9v.lib
  for each cell $i \in CellList$ **do**
    $\tau_{sub-V_{th}}(i)$= worst delay of all timing arcs from 0p4v.lib
  for each cell $i \in CellList$ **do**
    $DF(i)=\frac{\tau_{sub-V_{th}}(i)}{\tau_{super-V_{th}}(i)}$
  Calculate Median ($M_{current}$) and Quartiles (Q_1, Q_3) for DF
  $IRQ_{current} = Q_{3-current} - Q_{1-current}$
  $M_{old} = 0$
  while $|M_{current} - M_{old}| \leqslant 0.5$ **do**
    for each cell $i \in CellList$ **do**
      if $DF(i) \geqslant Q_3 + IRQ \times k$ **then** ▷ $0 \leqslant k < 1$
        $PruneList = Cell(i)$
      else
        $CellList = Cell(i)$
    $M_{old} = M_{current}$
    Calculate new Median ($M_{current}$) and Quartiles ($Q_1,\ Q_3$) for DF

ized at 0.4 V (SS, -40°C)and 0.9 V (TT, 25°C) in the typical corner. In this work, we use the NLDM model based library. Firstly, for all cells, the rise and fall delays of the worst timing arc are extracted from the center of the timing table in the liberty file using a custom script. Secondly, we define, a *degradation factor* (DF) for each cell, which is the ratio of cell delays between 0.4 V and 0.9 V. This degradation factor shows the rate at which the delay of a cell deteriorates with voltage scaling down to 0.4 V. The higher the degradation factor the worse the cell is. Note that a lower degradation factor does not mean that the cell is fast. The criteria for pruning are based on selecting cells with tightened degradation factors. The criteria for removing a cell is if its degradation factor is greater than the sum of 3$^{\text{rd}}$ quartile range and $k\times$ of interquartile range. k is a scaling factor between 0 and 1. In this work, we choose k to be 0.5. If k is 1 fewer cells are pruned which means we allow more variation among different types of cells, and if k= 0 more cells are pruned resulting in tighter criteria. The advantage or disadvantage of varying k is difficult to analyze. Essentially, this approach eliminates cells that do not have a homogeneous voltage-delay trend. The pruning keeps the cells that are, statistically speaking, below a given interquartile

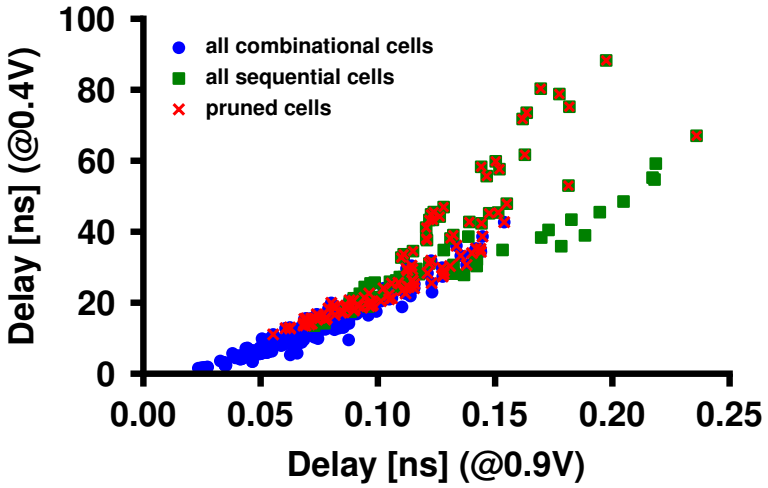


Figure 3.11. Correlation of propagation delay of standard cells at 0.9 V and 0.4 V.

threshold, which is a measure of dispersion equal to the difference between the 75^{th} and 25^{th} percentile. The delay distribution of all combinational cells, all sequential cells, and pruned cells at 0.4 V vs 0.9 V are shown in Fig. 3.11. We eliminate the outlier cells from the higher side of the distribution by our selection method as shown in Fig. 3.12. The remaining cells are again checked for outliers. This iteration is performed three times, as shown in Fig. 3.12. The average delay of the combinational and sequential cells after pruning reduces by 28% and 26%, respectively. The remaining cells are delay-homogeneous with voltage scaling. The pruned cells also match the general guidelines mentioned above. The sequential cells pruning is performed separately using the same method. In the standard cell library, there are 288 combinational cells and 92 sequential cells. The number of combinational cells remaining after pruning is 224 and the number of sequential cells remaining is 55. The total number of cells filtered is 101 out of 380 cells.

Note that, this methodology doesn't consider the global or local process variations of the cells in account. However, the proposed methodology can be extended for pruning cells based on the spread of delay due to global process variations. In the sub-V_{th} region of operation, the impact of process variations on delay are significantly high. Therefore, a degradation factor can be determined using a slow-slow corner liberty file and fast-fast corners liberty file for pruning based on the delay spread due to global process variations.

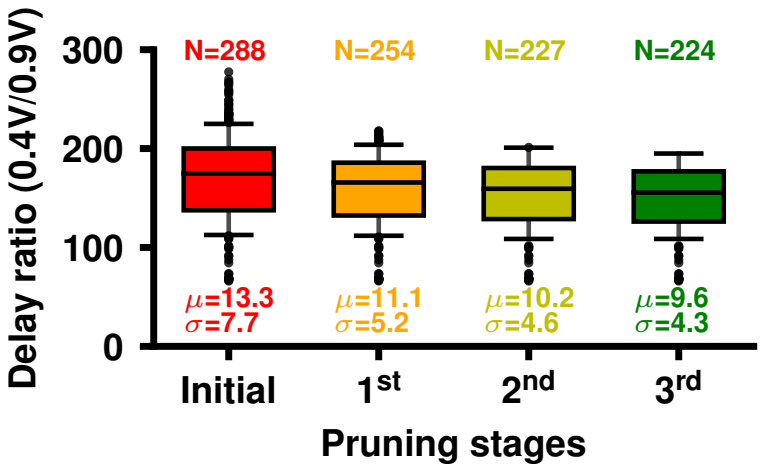


Figure 3.12. Box plot of library cells (N) after each pruning step. The average delay spread of all the cells decreases with every step.

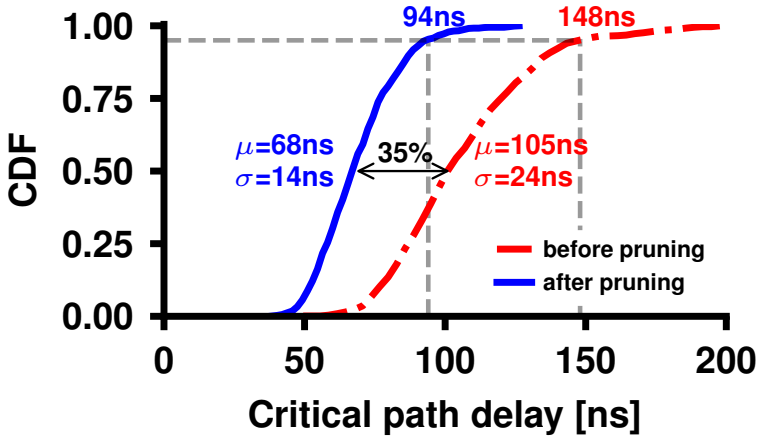


Figure 3.13. Variation of the critical path delay of the b18 benchmark with process after Monte Carlo simulations.

As per the generic guidelines, all 2-input cells should be allowed from synthesis. The list of 2-input and 3-input combinational cells filtered using our method

are shown in Table 3.3. Table 3.4 also shows the list of sequential cells which are filtered by our methodology. The filtered sequential cells are mostly scan flip-flops. The pruned cells also match the general guidelines mentioned above. Through our method, we could also determine other cells which are not filtered by the guidelines but degrade significantly as shown in the list of Fig. 3.11.

To demonstrate the impact of cell pruning, the b18 design (ITC benchmark) is synthesized at 0.4 V using the complete and pruned library for a target frequency of 7.5 MHz. The Monte Carlo simulation results of the critical paths of two syntheses are shown in Fig. 3.13. At the 95% yield point, the timing improvement for the design synthesized using the pruned library is 36% better as compared to the complete library. Additionally, the spread (σ/μ) of the delay also reduces by ∼15%. The trade-off is that after pruning the design area increases by 16%, due to the elimination of complex high-fan-in cells.

3.7 Summary

The standard cell library is one of the most important building blocks for all digital designs. Consequently, the design and choice of the low power and robust library cells become very relevant. We focused on practical techniques to handle the complexity of low voltage standard cell design and pruning, which are essential in any near/sub-V_{th} IC design. In this chapter, a new standard cell library is designed which can operate reliably in the near/sub-V_{th} region. The designed library is compatible with the existing foundry-provided standard cell library. The synthesis benchmark result of the mixed library shows that the newly designed cells are preferred over existing cells and the overall design achieves up to 50% leakage power reduction without any frequency penalty and <1% area overhead. Additionally, we propose a methodology for library cell pruning to detect and remove the cells from synthesis for near/sub-V_{th} operation. Compared to the generic guidelines for library pruning from the literature, the proposed technique detects bad cells based on the relative delay degradation with voltage scaling. The proposed library cell pruning methodology leads to delay spread reduction by 15% with process variations.

Table 3.3. The 2-input and 3-input cells which are pruned by the proposed filtering method.

2-input cells	**Strength**	**3-input cells**	**Strength**
AND4	X4	AND3X17	X17, X25, X33
AO22	X8	AO112X8	X8, X17, X33
AOI22	X4	AO212X8	X8
HA1	X8, X33	AO222X4	X4, X8, X17, X33
MUXI21	X3, X5	AOI112X5	X5
NAND2	X67	AOI13X5	X5
NAND4	X8, X17, X25, X33	AOI211X4	X4
NOR2A	X3	AOI222X4	X4
NOR4	X8, X17, X25, X33	CBI4I6X5	X5, X17, X25, X33
OA22	X8	FA1X8	X8, X33
OAI21	X5	MUX41X31	X8, X33
OAI22	X5	MX41X27	X7, X27
OR4	X20	NAND3AX18	X6, X18
		NAND3X18	X4, X9, X18
		NAND4ABX12	X6, X12, X18, X24
		NOR3X4	X4
		NOR4ABX6	X6
		OA112X17	X8, X17, X25, X33
		OA222X17	X8, X17, X33
		OAI112X10	X5, X10
		OAI211X10	X5, X10, X15
		OAI222X3	X3, X9
		XNOR2X17	X17
		XNOR3X16	X4, X8, X16, X25
		XOR3X17	X17, X24

Table 3.4. The sequential cells which are pruned by the proposed filtering method.

Sequential cells	**Strength**
DFPRQN	X17, X30
DFPRQ	X17
SDFPHRQN	X17, X30
SDFPHRQ	X8, X17, X30
SDFPQN	X8
SDFPQ	X17, X33
SDFPRQNT	X8, X17, X33
SDFPRQN	X17, X33
SDFPRQT	X8
SDFPRQ	X8, X17, X33
SDFPRSQT	X8, X17, X33
SDFPHRQN	X4
SDFPHRQ	X4
SDFPQNT	X4
SDFPQN	X4
SDFPQT	X4
SDFPQ	X4
SDFPRQNT	X4
SDFPRQN	X4
SDFPRQT	X4
SDFPRQ	X4
SDFPSQNT	X4
SDFPSQN	X4
SDFPSQT	X4
SDFPSQ	X4

Chapter 4

Charge recycling by voltage stacking

Low power embedded applications, such as IoT, wearable devices, and biomedical sensors are becoming increasingly computationally demanding, as more and more near-sensor data analysis is performed on-chip. The requirements for such platforms are to achieve low energy consumption and relatively high computation efficiency while meeting timing constraints. Voltage scaling to the near/sub-V_{th} region is a commonly used strategy to reduce power/energy consumption in ICs. Near/sub-V_{th} operation is challenging from both design and throughput perspectives, especially for always-on battery-powered applications. The reduced throughput is typically compensated by either adding an accelerator or by using parallel processing [51], [61], [64], [68], [138]. By way of example, always-on biomedical applications (e.g. EEG or ECG monitoring) require ultra-low power processors. But, the wide variety of complex algorithms demands dedicated hardware accelerators or multi-cores to meet the performance demands. As a consequence, voltage scaling is used to balance throughput and energy efficiency requirements.

Due to the constrained voltage scaling of foundry SRAMs, the near/sub-V_{th} operating systems require multiple voltage supplies or on-chip voltage regulators as shown in Fig. 1.5. The required distinct supply voltages result in high power conversion losses. Moreover, state-of-the-art on-chip voltage converters for near/sub-V_{th} have significant conversion losses and area overhead [47], [51], [64], [89], [92], [138].

Voltage-stacking of power domains for charge recycling is a promising method to reduce conversion losses. Voltage stacking is based on Kirchhoff's voltage law for series-connected power domains such that the ground of one domain becomes the power connection for the next. Thus, the domains are connected in a series stack for power delivery with all of them sharing the same current, and hence the charge is recycled [111]. Observe, however, that in voltage stacked systems the stacked power domains do not consume the same amount of current. Therefore, a voltage regulator is needed to stabilize the intermediate rail between the series-connected power domains [111], [139]–[141]. Balancing the current consumption mismatch of

voltage stacks is challenging when the stacks operate at super-V_{th} voltages because of the large current swings due to workload activity. Fortunately, voltage stacking in the near/sub-V_{th} region exhibits an almost constant leakage power consumption dominating the dynamic power. Therefore, the idea of voltage stacking for near/sub-V_{th} voltage becomes more feasible.

In this chapter, a focused review of the ultra-low-power designs with integrated power delivery and voltage stacked designs is presented, which serves as the basis for comparison. Furthermore, the motivation for the choice of architecture for the BrainWave application is presented. In this chapter, the design of the voltage stacked system, the controllers to balance the voltage stacks, level-shifters, and implementation of the chip for near/sub-V_{th} operation are presented in detail.

4.1 Background and related work

In Chapter 1, a detailed overview of the ultra-low-power near/sub-V_{th} chips are provided. In this section, we delve deep into the key specifications of the recent low-energy SoC with on-chip voltage regulators designed to operate in near/sub-V_{th}. An overview of the state-of-the-art voltage stacking implementations is also provided in this section.

4.1.1 Energy efficient near/sub-V_{th} region operation

Near/sub-V_{th} operation enables low energy consumption while achieving relatively high computation efficiency. Significant research towards minimizing energy consumption has been done in the past years [47], [49], [51], [59], [61], [64], [89], [138]. In [64], a RISC-V core along with a vector co-processor operating in near/sub-V_{th} region was demonstrated. Custom-designed 8T-SRAM cells instead of foundry delivered SRAMs are used for voltage scaling together with the logic circuit. The system achieves an energy consumption of 60 pJ/cycle with an energy efficiency of up to ∼20.9 MMACs/mW (million multiply and accumulate operations per second per milliwatt). In [61], a multi-core RISC-V implemented using a mixture of 8T and 6T based SRAMs for near/sub-V_{th} operation was presented. The multi-core system achieves an energy consumption of up to 20.7 pJ/cycle with a system efficiency of up to 95 MMACs/mW. However, the power delivery circuitry was not on-chip. In [89], a 32-bit lattice CPU was designed to operate in near-V_{th} along with a 3:1 on-chip DC-DC converter consuming 8 pJ/cycle. However, this work does not account for the SRAM supply voltage. An ultra-low-voltage system comprising an ARM-CM0+, 8 kB 10T-SRAM, 16 KB SRAM, and an AES-128 accelerator was shown in [49]. The complete system consumes as low as ∼23 pJ/cycle with a maximum of 53 pJ/cycle

while operating at 0.48 V. An MSP430 processor with 18 kB SRAMs and an energy consumption of 7 pJ/cycle was presented in [47]. The SRAM circuit operates at 1.0 V whereas the logic circuits were scaled down to 0.4 V.

Overall, the key factors impacting energy consumption are technology, processor architecture, multi-core processors, operating conditions, power conversion efficiency, and memory type and size. The above cited works have in common high computation efficiency and low energy consumption. A flexible system exploiting the benefits of near-V_{th} operation and parallel computing over multiple cores or accelerators is required for current embedded applications.

4.1.2 Power delivery for near/sub-V_{th} region operation

Voltage scaling to near/sub-V_{th} is the preferred choice to minimize energy consumption. Yet, even though the supply voltage of the logic can scale down to near/sub-V_{th}, note that the supply voltage of foundry SRAMs can hardly scale down. Therefore, such systems require multiple voltage converters. The employed voltage converters bring in extra area and power consumption overhead. Naturally, high energy efficiency and minimum area are often rigid requirements for the power delivery circuitry. To cope with this challenge, various types of on-chip voltage regulators are used such as LDOs and SCVRs [49]. The achieved power delivery efficiency highly depends on the conversion ratio (V_{out}/V_{in}). In the state of the arts, if the conversion ratio is <1/3, it is unlikely for the power delivery to achieve efficiency >80% [47], [49], [51], [89], [92]. In Chapter 1, a detailed survey of SoC with integrated power delivery circuit is provided. From the surveyed papers it is evident that for near/sub-V_{th} the achieved system efficiency is in the range of 70-85% with significant area overhead. In the literature, the area overhead for a microprocessor based system including on-chip power delivery is up to 38% [49], [51], [64], [89], [138]. Note that in all previous works, one SCVR is used in the chip in addition to an LDO. The use of two SCVR can result in significant area overhead. Therefore, in this work, one SCVR followed by an LDO is used for comparison. In many cases, the designs become complex due to the multiple supply voltages required for IO-cells and core.

Voltage-stacked systems: All state-of-the-art implementations of voltage-stacked systems are for the above-V_{th} region and consist of either simple circuit blocks or disconnected multi-core systems [111], [139]–[141]. In all prior voltage-stacked systems, on-chip SCVRs are used to balance the intermediate rails between the stacks, which brings significant area overhead (up to 36% of the chip area) [111], [139]–[141]. In [141], a multiple CPU system, with each CPU operating at 1.2 V, is used to demonstrate the voltage stacking gains, achieving a total system efficiency of 87.1% while

using several simple LDOs with a maximum conversion efficiency of 44.4%. The area overhead of the on-chip LDOs is ∼3%. However, the system requires additional off-chip tank capacitors of at least 1 μF for charge recycling.

Although voltage stacking is a known technique, the previous works suffer from the large variation of the intermediate rail due to the unbalanced activity of the system. In fact, the control circuits become increasingly complex for balancing the stacks in high performance designs [140]. In [140], application-level voltage smoothing techniques are used to mitigate the stack balancing issue, finally achieving up to 93% system efficiency. Despite these advances, the state-of-the-art voltage-stacked systems cannot be simply scaled to operate at near/sub-V_{th} voltages. Usually, the middle node between power domains is fixed and is also biased above the threshold voltage to meet the SRAM supply voltage requirement. In other words, in a flat system, the SRAM and logic can have different supplies so that they do not depend on each other, whereas in the voltage stacked system, voltage scaling cannot be done independently without affecting the supply requirements of the series connected domains. In [142], an above-V_{th} operation microcontroller system was implemented in a two-level voltage-stacked fashion with SRAMs in the top stack and logic in the bottom stack. The balancing of the voltage stack is accomplished by using an SCVR, achieving up to 96% system efficiency with an area overhead of 33%. However, since the middle voltage node is optimized at half of the stack voltage, the efficiency of this SCVR is sub-optimal if the two stack voltages are different. In this case, the logic circuit cannot be scaled to the near/sub-V_{th} region because of the high voltage required for the SRAM blocks.

The proposed voltage-stacked system in this chapter is designed to operate at a supply voltage of 1.8 V$\pm$5% eliminating the requirement of multiple supplies for IO-cells and core. The balancing of the intermediate voltage rails between the power domains is achieved by dedicated controllers which results in minimal area overhead, compared to an on-chip LDO or SCVR. Since the logic circuit is operating in the near/sub-V_{th} region, the controller design complexity is significantly reduced.

4.2 BrainWave processing platform

Biomedical signal processing platforms are commonly designed with multiple processor cores and are typically coupled with hardware accelerators [16], [22]–[24]. Unfortunately, these architectures either lack energy efficiency if the architecture is fully programmable or are specialized towards a limited set of kernels. For emerging and complex monitoring tasks such as non-convulsive epileptic seizure detection and Parkinson's Disease FoG prediction, research is ongoing on what algorithms and sensors work best. These applications demand an energy-efficient and flexible platform.

Kwong *et al.* [22] employ a micro-processor with hardware accelerators for common bio-medical kernels (FFT, CORDIC, FIR, and Median filtering) and report platform-level energy-savings over 10× on two biomedical applications over a processor-only mapping. Lee *et al.* [23] propose a more flexible approach sharing a CORDIC, specialized data-path unit and a scratch-pad memory between an SVM and active-learning accelerator. This solution results in a 68.3× speedup, and 144.7× energy reduction with respect to a processor-only approach. More recently, Coarse-Grained Reconfigurable Architectures (CGRAs) are being advertised as a good compromise between flexibility and energy efficiency [62], [143], [144]. Das *et al.* [144] introduce a CGRA as a co-processor of a multi-core platform targeted towards ultra-low power edge processing. They obtain an energy gain of 6–18× for several common signal processing kernels, compared to a RISC processor. The authors of [143] extend a multi-core system with a CGRA and report 37.2% energy savings over a multi-core-only implementation on a complex ECG algorithm. In this work, we consider a signal processing system with CGRA running a more complex EEG-based seizure detection algorithm.

The BrainWave processor is a processor platform that is flexible and capable of performing energy-efficient signal processing. The BrainWave processor is an always-on processor which offloads complex EEG features to the CGRA and exploits near/sub-V_{th} computing to improve energy efficiency. The BrainWave processor architecture is depicted in Fig. 4.1. The seizure detection algorithm runs on the single-issue RISC-V core [145] with a tightly-coupled program (IMEM) and data memories (DMEM). The RISC-V core can tell the loader to start loading a new kernel and network configuration. When loading is completed, the CGRA notifies the RISC-V. Then the RISC-V can issue the execution of a kernel. While the CGRA is processing, the RISC-V can either process something else, or goes into idle mode and waits for the program to complete. The DMEM is sized to store up to 20 channels × 256 samples/epoch × 2 byte/sample elements (× 2 for double-buffering) and some scratchpad memory to perform the feature computations. An UART is included to interface with an external radio module to notify a medical expert in case of emergency.

This CGRA enables flexible and energy-efficient processing by providing programmable function units (FUs) and a reconfigurable data-path to bypass the register file [146]. These FUs operate in lock-step and act as a VLIW processor. Vector-processing (SIMD) is naturally supported since multiple FUs can share the same instructions and data via the reconfigurable data and instruction network. The CGRA has a private memory where its network configurations and programs are stored. The CGRA programs and configurations can be reused by consecutive acceleration requests to reduce reconfiguration overhead. Typically, the RISC-V core issues a new

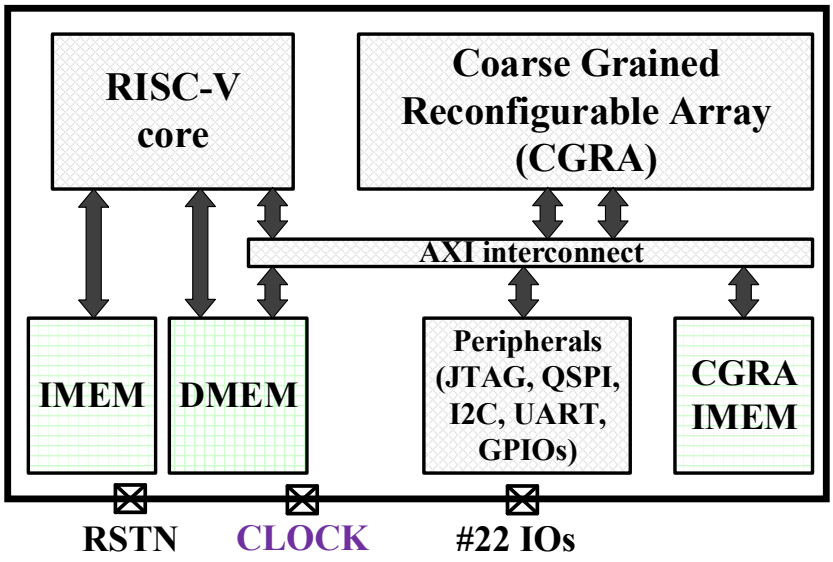


Figure 4.1. BrainWave processor architecture.

acceleration request. After this request, the CGRA starts executing the preloaded program. Important parameters such as which kernel should be executed and where the data is stored, are read from a fixed location in the (shared) DMEM.

The internal structure of the instantiated CGRA is illustrated in Fig. 4.2. The CGRA consists of 6 different types of FUs, which can perform RISC-like instructions. More information on the instruction set can be found in [146]. Small local standard-cell memories (SCM) are used for local processing. Every Load-Store Unit (LSU) can access the shared data memory to access the EEG data. The currently loaded program is also stored in SCMs called local memories (LMs). The instantiated CGRA contains 20 FUs and 11 instruction decoders (IF/ID) that can be connected to one or more FUs. The CGRA supports up to 4 multiply-accumulations per cycle using a 4-wide SIMD operation and can run the complex kernels 6–10× faster as compared to the RISC-V core, resulting in improved energy efficiency [146]. The CGRA contains 20 functional units which can perform up to four 32-bit multiply-accumulations, eight 32-bit integer operations, and four memory operations per cycle.

The performance required for EEG seizure detection and classification algorithms on our platform is ∼2.5 MHz. The application of seizure detection and classification

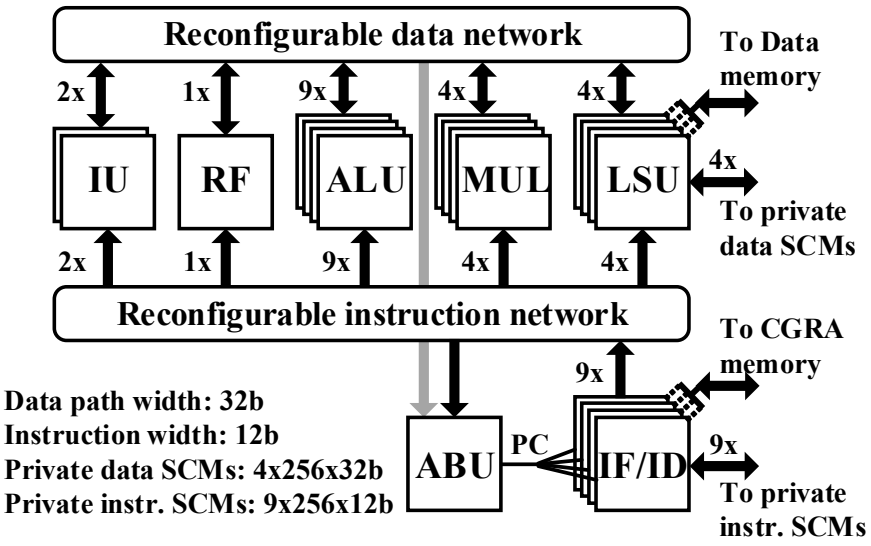


Figure 4.2. Instantiated CGRA with interfaces.

demands always-on operation to monitor the seizure and to raise an alarm on detection. The CGRA is sized to run complex kernels such as matrix multiplication (MatMul), Butterworth filtering, approximate entropy, FFT, wavelet decomposition, and sorting. The CGRA enables parallel and energy-efficient processing for biomedical signal processing applications while operating in the near/sub-V_{th} region.

4.3 Voltage stacking for near/sub-V_{th} region operation: Design enablement

The BrainWave processing system consisting of a RISC-V core (25 k gates) [145], a reconfigurable and programmable energy-efficient accelerator CGRA (330 k gates) [146], JTAG for programming, and peripherals (15 k gates) including 17-GPIO's muxed with QSPI, I^2C, and UART is partitioned for voltage stacking. The system consists of 32 kB program memory, 32 kB data memory, and a 16 kB CGRA dedicated memory. A total of 10 kB local data/instruction memories for the CGRA are implemented with latches to enable voltage scaling. When the CGRA is active it operates from its local memories, which minimizes SRAM activity in the TOP stack.

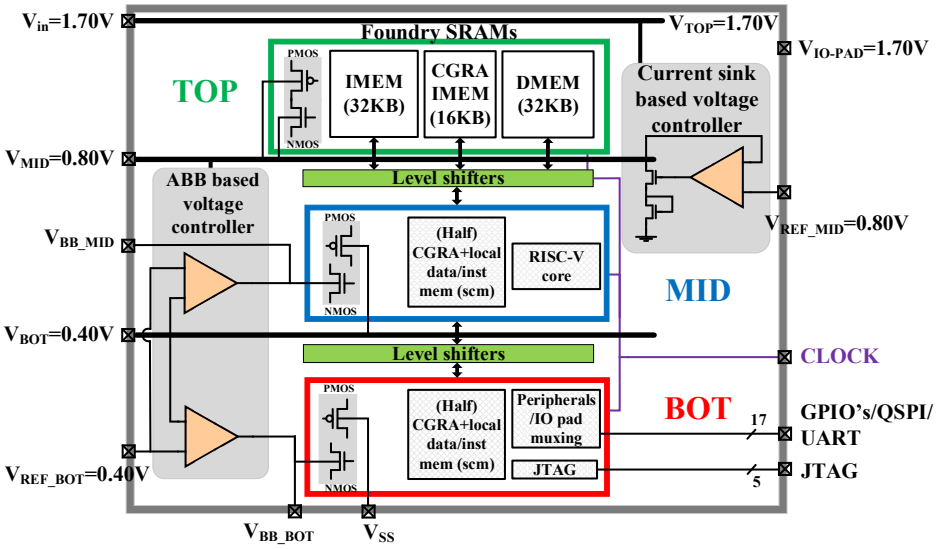


Figure 4.3. System partitioning for three-level voltage stacking.

Without loss of generality, in the remaining of the chapter, we operate the voltage stacked system at the worst case supply voltage of 1.7 V (1.8 V−5%). The worst case IO-pad supply voltage is also 1.7 V. For the voltage stacking implementation, the BrainWave processor architecture has to be partitioned in stacks in such a way that the stack balancing is simpler. Additionally, the partitioning has to meet the voltage requirements of SRAMs in the design. Furthermore, the logic circuit should operate in the near/sub-V_{th} region for low energy consumption. For simpler input and output level-shifting of signals between logic and IO-cells, the peripherals are placed in the bottom stack, operating at 0.4 V. The CGRA is a regular block which is partitioned in two equal parts with equal power consumption, while minimizing the number of required level-shifters. The number of gates in the RISC-V core is almost the same as of the peripherals. Therefore, RISC-V is placed in the middle stack. This leads to placing the SRAMs in the top stack operating at 0.9 V.

The three-layer partitioned voltage-stacked system is shown in Fig. 4.3. The top stack consists of 80 kB SRAMs operating at 0.9 V (1.7 V↔0.8 V). The middle stack is composed of the RISC-V core and half of the CGRA, operating at 0.4 V (0.8 V↔0.4 V). The bottom stack contains the remaining half of the CGRA, peripherals, and IO-pad control logic, operating at 0.4 V (0.4 V↔0 V). There are other partitioning options possible, but the used partitioning strategy is well-balanced in

Table 4.1. Synthesis result using the LVT and RVT standard-cells with the same timing constraints.

	Number of cells	**Area (mm^2)**	**Leakage Logic (μW)**	**Leakage SRAMs (μW)**	**Total power (μW)**
LVT	350k	0.68	35.6	57.6	168
RVT	422k (+21%)	0.96 (+41%)	36.8 (+3%)	57.6	177 (+5%)

terms of current consumption of the stack and simple from the stack balancing point of view.

The foundry 28-nm FDSOI was used for the implementation of the voltage stacked system. The use of 28-nm FDSOI brings additional advantages over bulk CMOS technology for voltage stacking. The 28-nm FDSOI provides the option of a triple-well. The triple-well is very useful in voltage stacking to isolate the body of the power domains with raised ground voltage of transistors. In bulk CMOS technologies without the option of triple-well, voltage stacking technique suffers from body-effect. Additionally, 28-nm FDSOI is latch-up safe as all the wells are isolated by buried oxide (BOX). In addition, the 28-nm FDSOI provide RVT and LVT transistor flavors. The foundry 28-nm FDSOI high density 8-track LVT standard-cell library was used for the implementation. There are multiple reasons to use LVT cells instead of RVT cells. The RVT cells are better suited for duty-cycled applications where leakage in the idle mode is of utmost importance. Our application is for always-on usage where dynamic power is more important. Additionally, there are multiple timing constraints in our platform. The most critical one is in between the combinational switch-box network in the CGRA. The LVT cells could meet this timing constraint (75 ns) whereas the RVT cells could not, with a timing violation of 45 ns. Even with the timing violation, the synthesis results with RVT cells are shown in Table 4.1. Synthesis using RVT cells would have resulted in ∼21% more logic cells as well as 41% larger area as compared to the synthesis using LVT cells due to less buffering/driving strength required in the near/sub-threshold region. But please recall that RVT cells don't meet our timing constraints. Furthermore, LVT cells can robustly scale down to 0.4 V with sufficiently high performance as compared to RVT cells. Moreover, voltage scaling down to 0.4 V is required in our voltage stacking scenario as the two logic stacks operate at 0.4 V and the SRAM operates at 0.9 V. This fits well within the worst-case limit of the industrial standard supply voltage of 1.7 V (1.8 V±5%).

The partitioning of logic between stacks is done based on a post-synthesis leakage power distribution. The leakage power breakdown of the system before partitioning is shown in Fig. 4.4a. The power breakdown shows that the SRAMs consume 59%

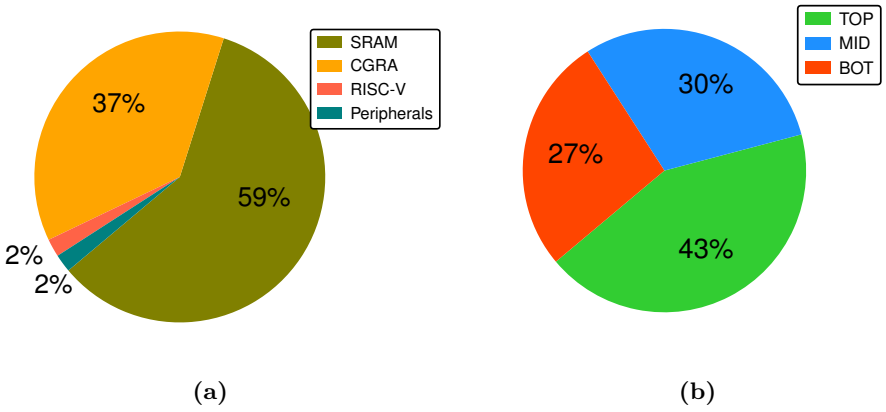


Figure 4.4. (a) Leakage power breakdown of the BrainWave processor. (b) Leakage current breakdown after system partitioning among stacks.

of the total leakage power consumption. For voltage stacking the current flowing through the partitioned stacks is of more interest, shown in Fig. 4.4b. The leakage current breakdown shows that the current consumption of both the MID and BOT stacks is almost the same and that it is 13% lower as compared to the TOP stack.

Special ultra-low-power low-to-high and high-to-low level-shifters were designed for the logic signals between the voltage stacks. The designed level-shifters are standard-cell compatible and operate over a wide voltage range and can also operate in flat mode when the voltage domains are connected in parallel. A simple current sink (CS) voltage controller was designed to control the voltage between the TOP and MID stacks at 0.8 V. The balancing of the intermediate rail between the MID and BOT stacks at 0.4 V is achieved by means of an ABB controller. The CGRA partitioning between the MID and BOT stacks is done based on an equal distribution of logic units. The LM, LSU, ALU, MUL, and other FUs are equally partitioned between the MID and BOT stacks as shown in Fig. 4.5. The multipliers are the most power consuming block among all the blocks whereas the LMs are the standard cell based memories with the majority of gates.

4.3.1 Design of the current sink based voltage controller

In state-of-the-art voltage stacked systems the voltage regulation of the intermediate rail (between the stacks) is done through an LDO or an SCVR [111], [139]–[142]. In this work, the total leakage current flowing from the TOP stack is higher than the MID/BOT stack current consumption as shown in Fig. 4.4. Note that the in-

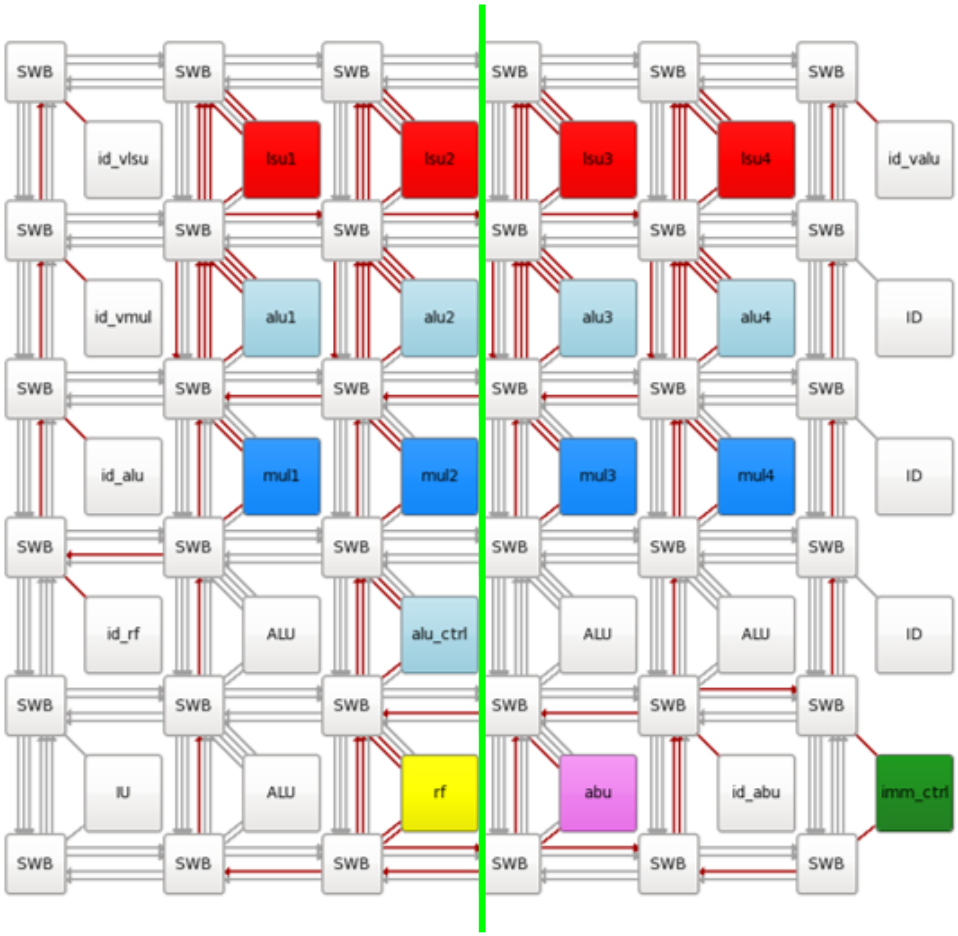


Figure 4.5. CGRA partitioning strategy among the MID and BOT stacks.

termediate rail voltage changes due to the current difference between the TOP and MID stacks. Therefore, the excessive current from the TOP stack is sunk by the CS controller.

The CS controller, shown in Fig. 4.6, compares the voltage of the intermediate rail (V_{MID}) with an input reference voltage (V_{REF_MID}=0.8 V) and adjusts the gate voltage of the current sink transistors to sink the excessive current while regulating the voltage of V_{MID} at 0.8 V. The intermediate V_{MID} rail is the input to the voltage amplifier (M6) and connected to transistors (P1, P2, and P3) which are responsible for the current sinking. The reference voltage (V_{REF_MID}) is connected to the other

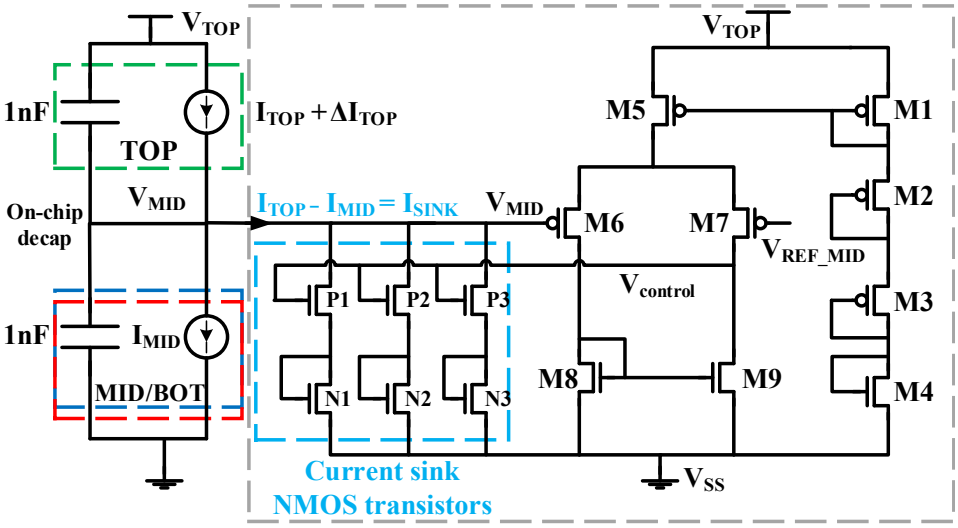


Figure 4.6. CS voltage controller schematic and test-bench.

input (M7) of the amplifier. Transistors M1-M4 constitute the biasing circuit for the amplifier. The output of the amplifier is controlling the gate voltage of the current sink transistors. The diode-connected transistors (N1-N3) in the current sink path are used to raise the bias voltage of the V_{MID} to 0.8 V. These transistors small signal model is a resistance. Two CS controllers are used on both sides of the chip to sink the excessive current uniformly. Note that, the CS controller operates at a supply voltage of 1.7 V, hence the circuit is designed using thick-oxide transistors. The reference voltage generation can be achieved in several ways and falls outside the scope of the thesis. The design of voltage reference circuitry is widely explored in literature and also consumes a negligible current [147].

Small-signal circuit modelling: The CS controller small signal model in steady state is shown in Fig. 4.7. The SRAM and logic circuit power domains are modeled by small signal leakage resistance (r_{top}, r_{midbot}). The amplifier is modelled by a voltage dependent voltage source of gain A with an output impedance of r_o. The load for the amplifier is the gate capacitance of the current sink transistors. The loop gain for the small signal model is given by

$$A_{loop} = \frac{A g_{m1} g_{m2} (r_{top} || r_{midbot})}{g_{m1} + g_{m2}}. \tag{4.1}$$

The simulated voltage gain of the designed CS amplifier is ∼32. The small signal

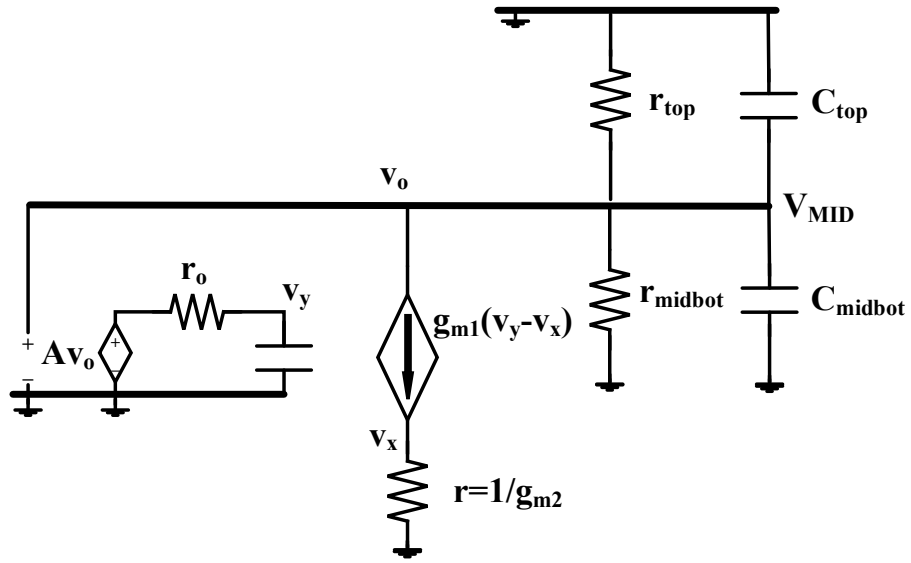
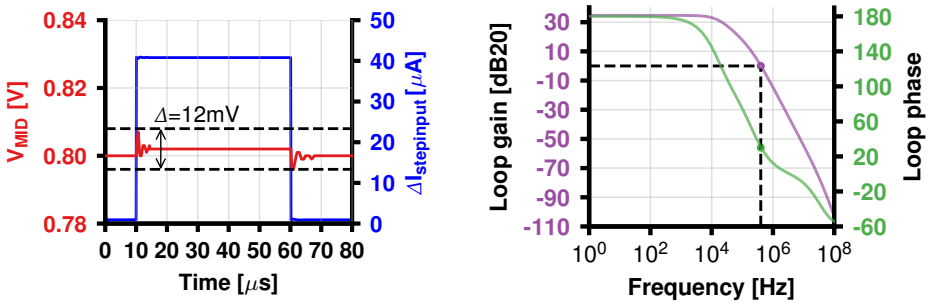


Figure 4.7. CS voltage controller small signal model.



(a) (b)

Figure 4.8. Simulation results of the CS controller. (a) The simulated voltage regulation of V_{MID} rail by the CS controller upon forcing a step current and (b) closed loop stability simulation showing the loop gain and phase for the CS controller.

model parameters for the circuit are estimated by simulation.

The simulation of the proposed CS controller was performed using the circuit

model shown in Fig. 4.6. The voltage stacked system is modelled by current sources and on-chip decoupling capacitors. The imbalance at the intermediate rail (V_{MID}) between the TOP and MID stacks is created by forcing a step current ($\Delta I_{stepinput}$). The transient response of the CS controller for a step current change of 40 μA is shown in Fig. 4.8a. The simulated peak to peak voltage overshoot of the intermediate V_{MID} rail is within $\pm$5% of 0.8 V with a response time of $\sim$400 ns. The outcome of the closed-loop stability simulation is shown in Fig. 4.8b. A closed-loop phase margin of 30° indicates that the closed-loop system is stable.

Failure analysis of voltage stack: The failure point of the system is when $I_{TOP} < I_{MID}$, since we have only used a current sink controller to balance the V_{MID} rail in the voltage stack. In our system, having I_{TOP} equal to I_{MID} is very unlikely as the SRAM's power consumption is dominant. A back of the envelope calculation of current consumption of the SRAMs and logic stacks, with different process and temperature combinations, using the available/characterized liberty files is performed for the proposed voltage stack. In our design, we have a total of 80 kB SRAM in the TOP stack. Please bear in mind that we have only a limited number of SRAM liberty files available. The current consumption of the SRAMs is shown in the Table 4.2. The leakage current consumption of the MID stack is calculated using the current consumption of an equivalent NAND2 gate. The number of equivalent NAND2 gates in the MID stack is 163k (total cell area/area of NAND2). The leakage current consumption at different process and temperature corners for the MID stack are shown in Table 4.2.

The analysis shows that the current consumption of SRAM is always greater than that of the logic stack across different process and temperature corners. In the corner SS, 0.90 V/-25°C, the leakage current consumption of the SRAMs is $\sim$10$\times$ higher than the MID stack current consumption. In the FF, 0.90 /125°C corner, the difference has reduced to $\sim$1.3$\times$ (extrapolated using data of FF, 0.95 /125°C and FF, 1.15 /125°C). But, please recall that we actually did timing closure for FF corner and 80°C. Thus, we expect the difference to be >1.3$\times$.

4.3.2 Design of the adaptive body bias based voltage controller

In the state-of-the-art voltage stacking implementations, the intermediate rail between the stacks is regulated by sourcing or sinking the current difference between the stacks using a voltage regulator [111], [139]–[142]. Despite that the MID and BOT stacks are partitioned almost equally, the balancing of the intermediate rail between the stacks is challenging due to the unpredictability of the switching activ-

Table 4.2. Process and temperature variation analysis of the current consumption by TOP and MID/BOT stacks in the voltage stacked system.

SRAMs in the TOP stack variation				**MID/BOT power domain variation**			
Voltage / Temperature	**SS (μA)**	**TT (μA)**	**FF (μA)**	**Voltage / Temperature**	**SS (μA)**	**TT (μA)**	**FF (μA)**
0.90V/-25°C	**9.9**			**0.40V/-25°C**	**0.9**	1.5	3.3
				0.40V/0°C	4.6	7.4	13.4
0.90V/25°C		**63.9**		**0.40V/25°C**	19.8	**30.1**	50.7
				0.40V/80°C	254.2	369.5	575.1
0.90V/125°C			**3530.5**	**0.40V/125°C**	1279.5	1815.8	**2715.9**
0.95V/125°C			4408.9				
1.15V/125°C			7893.1				

ity that is caused by workload variation. In this work, a body bias based voltage regulating controller is proposed to control the intermediate rail V_{BOT} at 0.4 V as shown in Fig. 4.9. The ABB controller is based on reducing/increasing the current consumption of the stacks by changing the body bias in opposite directions, instead of sinking or sourcing the current difference as is known in the literature.

The chip was designed using LVT standard cells. LVT NMOS and PMOS transistors are fabricated using a flip-well construction where NMOS transistors are placed in the N-well and PMOS transistors are placed in the P-well, respectively [123]. The LVT FBB range goes from 0 V to +3 V and from –3 V to 0 V for NMOS and PMOS transistors, respectively. The default body bias condition for LVT is when the N-well and P-well are connected to the lowest voltage (GND) of the power domain. The wide range of body biasing is sufficient to balance the stacks for near/sub-V_{th} operations.

In our design, we body biased only the NMOS transistors to balance the current consumption of the MID and BOT stacks. The ABB controller senses the voltage of the intermediate rail V_{BOT} between the MID and BOT stacks, compares it against V_{REF_BOT}, and adjusts the body bias voltage of the NMOS transistors in opposite directions in the corresponding stacks. The body of the PMOS transistors is connected to the ground of the corresponding power domain. Two different amplifiers were designed to provide the body bias voltage for the NMOS transistors in the MID and BOT stacks. The designed amplifiers are responsible for driving the body (NWELL) of the MID and BOT stacks. The load for body biasing is almost purely capacitive (∼600 pF) as shown in Fig. 4.10a. The schematic of the designed ABB controller is shown in Fig. 4.10b. Transistors M1-M3 are used to generate the

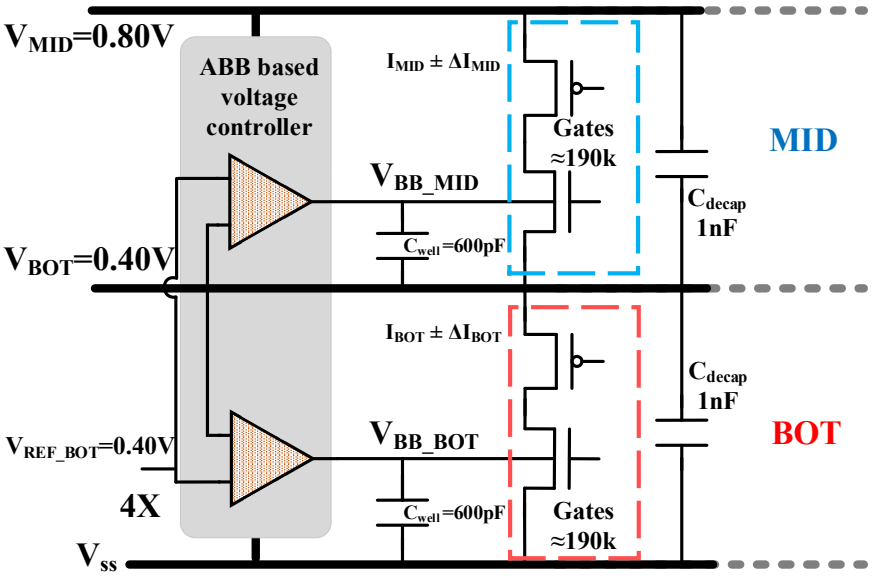


Figure 4.9. ABB controller operation.

bias voltage for M4, M12 of the two amplifiers by forming a current mirror circuit. Transistors M4-M11 and M12-18 constitute the designed two-stage amplifiers. V_{BB_MID} is the output of the second stage (M10, M11) of the amplifier. Similarly, V_{BB_BOT} is the output of the second stage (M17, M18) of the amplifier. We estimated the body capacitance from the design documents and some basic calculations: $C_{body} = C_{gb} + C_{bd} + C_{bs}$ ≈0.9 fF/ transistor. In our design, the number of cells in the MID/BOT power domains is ∼180k. We assumed three LVT NMOS transistors per cell on average. The calculated total body capacitance is 486pF (0.9 fF×3×180000). We assumed the body capacitance to be ∼600pF by considering some missed capacitances (decaps, parasitic, etc.). The estimated capacitance was verified against the data available in the literature [80]. At design time, the NMOS transistors in the MID and BOT stacks are FBB by 200 mV, resulting in V_{BB_MID}=600 mV and V_{BB_BOT}=200 mV. The ABB controller is capable of both FBB and RBB in the MID stack and only FBB in the BOT stack depending on the ratio of imbalance. The ABB controller behavior and output voltage ranges are shown in Table 4.3. On power-up, the ABB controller adjusts the NMOS transistor body bias voltage to balance the initial current difference between the MID and BOT stacks due to global PVT conditions. Subsequently, the current mismatch during runtime, due to differ-

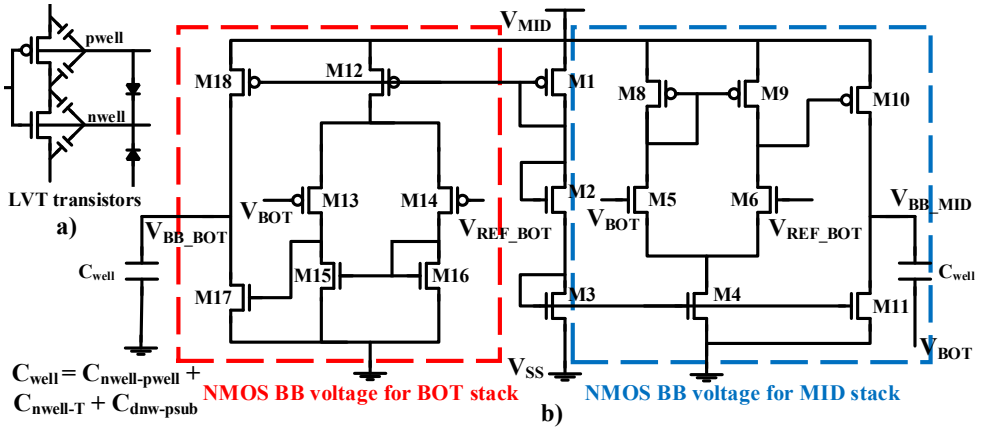


Figure 4.10. (a) Load-model for ABB controllers. (b) ABB controller circuit diagram.

Table 4.3. Characteristics of ABB controller and range of FBB/RBB on NMOS transistors in the MID and BOT stacks.

Condition	$\mathbf{V_{BB\_MID}}$	$\mathbf{V_{BB\_BOT}}$
Initial	600 mV	200 mV
$\mathbf{I_{MID} > I_{BOT}}$	Decrease	Increase
$\mathbf{I_{MID} < I_{BOT}}$	Increase	Decrease
Range of FBB	V_{BOT} to V_{MID}	V_{SS} to V_{MID}
Range of RBB	V_{BOT} to V_{SS}	NO

ent workloads, is dynamically balanced by the ABB controller. The ABB controller is designed to control the intermediate rail V_{BOT} to within ±5% of 0.4 V (<40 mV).

Small-signal circuit modelling: The ABB controller small signal model assuming no activity is shown in Fig. 4.11. The NMOS transistor is model by body-effect transconductance (g_{mb}) and small signal output resistance (r_{ds}). The amplifier is modelled by a voltage dependent voltage source of gain A with an output impedance of r_o. The load for the amplifier is the power domain's well capacitance (C_{well}). The small signal model is simplified by a symmetric half circuit model shown in Fig. 4.12. The loop

$$A_{loop} = A g_{mb} r_{ds}. \tag{4.2}$$

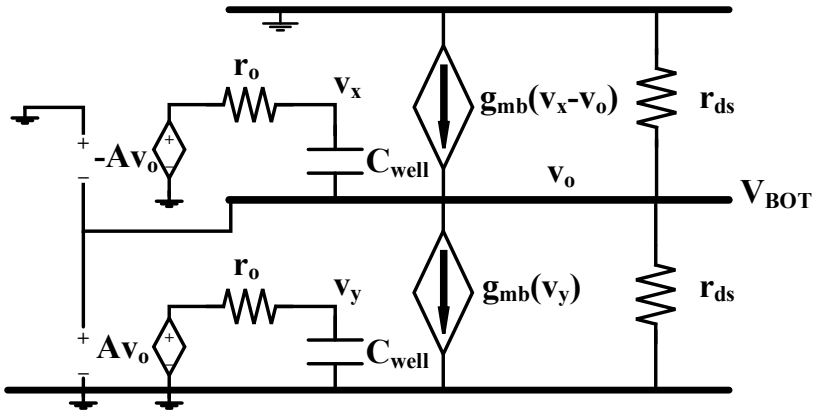


Figure 4.11. Small signal model for the ABB controller and the voltage stack.

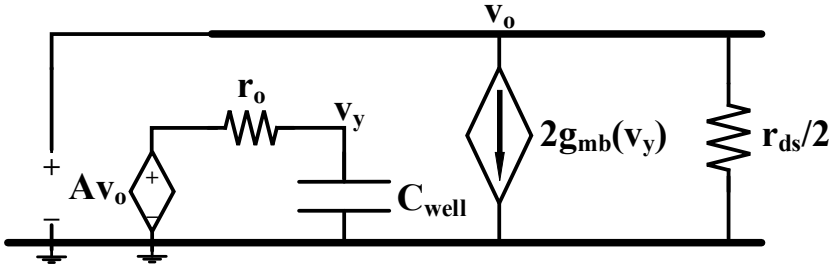
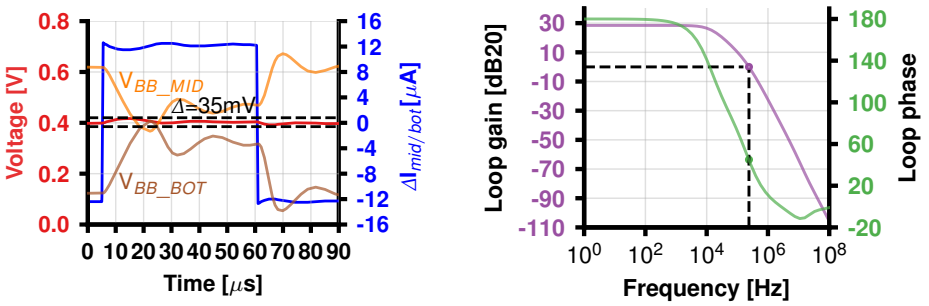


Figure 4.12. Simplified half circuit model for the closed loop ABB circuit.

The open loop gain of the system is given by

$$A_{loop}(s) = \frac{A_{loop}}{1 + s\tau}, \tau = r_o C_{well}. \tag{4.3}$$

The settling time for the controller while balancing the intermediate rail voltage to V_{REF} is determined by the dominant pole in the system. The well capacitance is estimated to be ∼600 pF. The simulated voltage gain of the designed amplifiers is ∼30. The small signal model parameters for the circuit are estimated by simulation. The small change in the current between the stacks (ΔI) produces a voltage deviation (ΔV) at the output node. The transfer function of the closed loop controller is given



(a) **(b)**

Figure 4.13. Simulation results of the ABB controllers. (a) The simulated voltage regulation of V_{BOT} rail by the ABB controller upon forcing a step current and (b) closed loop stability showing the loop gain and phase of the ABB controller.

by

$$\Delta v_o(s) = \frac{r_{ds}}{2(1 + A_{loop}(s))} \Delta i_{in}(s). \tag{4.4}$$

The ABB controller was simulated using the circuit model shown in Fig. 4.9. The circuit simulation was performed by creating an imbalance between the MID and BOT stacks using a current source ($\Delta I_{MID/BOT}$). The peak to peak voltage overshoot (35 mV) of the intermediate rail V_{BOT} is $< \pm 5\%$ of 0.4 V for a peak to peak current variation of 20 μA, see Fig. 4.13a. The response time for the step current change is $<1\,\mu s$. The ABB controller can stabilize the intermediate rail at an average voltage of 0.4 V. As shown in Fig. 4.13b, the simulated closed-loop phase margin is 40°, indicating that it is stable.

Impact of asymmetric body biasing: In our design, we apply FBB to the NMOS transistors while the PMOS transistors are connected to the ground of the power domain. The cells become faster with an increase in FBB. To further investigate the asymmetric body-bias, we simulated a NAND and a NOR cell. The gate fall delays improves by ~20% for an FBB increase from 0 to 400 mV. The SPICE simulation of a NAND and a NOR cell from the LVT library (PB10) are shown in Table 4.4 for different body bias voltages.

The noise margin degradation is not that significant as shown in the Table 4.5. The noise margin high (NM_H) improves by 8% and the noise margin low (NM_L) degrades by 10% when FBB goes from 0 to 400 mV. This shows that the overall noise

Table 4.4. Impact on cells delays due to NMOS FBB. SPICE simulation of a NAND and a NOR cell from PB10 library operating at 0.40 V, typical corner, and 25°C.

Supply	PMOS BB	NMOS BB	Fall delay	Rise delay	Fall delay	Rise delay
0.4V	0 (default)	0 (default)	328ps	408ps	182ps	1294ps
0.4V	0 (default)	200mV	284ps	405ps	166ps	1279ps
0.4V	0 (default)	400mV	254ps	402ps	150ps	1263ps

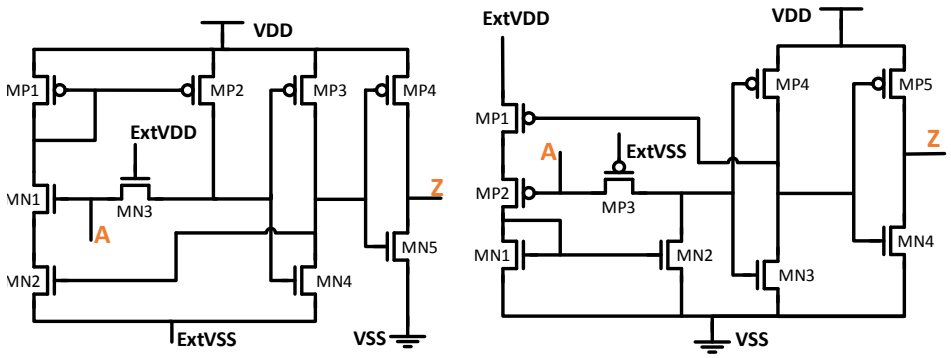
Table 4.5. Noise margin variation due to NMOS FBB. SPICE simulation of a NAND cell from PB10 library operating at 0.40 V, typical corner, and 25°C.

Supply	PMOS BB	NMOS BB	$NM_H=V_{OH}-V_{IH}$	$NM_L=V_{IL}-V_{OL}$
0.4V	0 (default)	0 (default)	0.154V (0.384V-0.230V)	0.164V (0.176V-0.012V)
0.4V	0 (default)	0.2V	0.163V (0.384V-0.221V)	0.156V (0.168V-0.012V)
0.4V	0 (default)	0.4V	0.170V (0.384V-0.214V)	0.148V (0.160V-0.012V)

margin degradation is not significant. Again, since we use LVT cells the degradation with process and temperature variations is better as compared to RVT cells. The SPICE simulation for noise margin of a NAND gate from the LVT library (PB10) is shown in Table 4.5 for different body bias voltages.

4.3.3 Design of level-shifters for voltage stacking

The three voltage stacks communicate via level-shifters. The level-shifters need to be standard cell compatible with minimum delay, power, and area overhead. In this three-level voltage stacking system, four types of level-shifters are required between adjacent power domains: level-shifters for BOT-to-MID and for MID-to-TOP. The low-to-high level-shifters are responsible for converting the logic level from (0 V-0.4 V) to (0.4 V-0.8 V) for the signals from the BOT to MID stacks. Likewise, high-to-low level-shifters are required to convert the logic levels from (0.8 V-0.4 V) to (0.4 V-0 V). These level-shifters were designed using thin-oxide transistors and are compatible with the 8-track standard cell library. The low-to-high level-shifters and high-to-low level-shifters between the TOP and MID stacks were designed to convert logic levels from (0.4 V-0.8 V) to (0.8 V-1.7 V) and (1.7 V-0.8 V) to (0.8 V-0.4 V), respectively. Worth observing is that a 1.3 V (which is the maximum voltage drop across the level-



(a) Low-to-high level-shifter. **(b)** High-to-low level-shifter.

Figure 4.14. The schematics of designed level-shifters.

shifter) is harmful to thin-oxide devices. Hence, in this work, thick-oxide transistors were used for the level-shifter design between the TOP and MID stacks to fulfil reliability requirements. The designed high-voltage level-shifters are double-height (16-track) and compatible with the standard logic cells. The designed low-to-high level-shifter is a modified version of the level-shifter in [148], which is improved to operate reliably in the voltage-stacked mode. The schematic of the low-to-high and high-to-low level-shifters are shown in Fig. 4.14. The same schematic is used for thin oxide and thick oxide transistor based level-shifters.

The operation of the low-to-high level-shifter in Fig. 4.14a is as follows. ExtVDD and ExtVSS are the input voltages connected to 0.4 V and 0 V, respectively. VDD and VSS are the voltages of the signals in the output power domain, which are connected to 0.8 V and 0.4 V, respectively. A pass transistor (MN3) is used for speeding up the fall transition. The feedback transistor (MN2) mitigates the high static current in the current mirror during the high output. When the input (A) goes high, MN1 is turned on to generate the mirror current through MP2. The mirror current charges the input of the inverter that is composed of MP3 and MN4. Then the output of this inverter is discharged through MN4, and the output (Z) goes high. The output of the inverter (composed of MP3 and MN4) turns off MN2 to cut the static current through MN1 and MP1. When the input goes low, MN1 is switched off to disable the mirror current. The pass transistor MN3 is turned on to quickly discharge the input node of the inverter composed of MP3 and MN4. Then, the output Z goes low.

The designed high-to-low level-shifter is a complementary version of the low-to-high level-shifter as shown in Fig. 4.14b. If the ExtVSS is connected to VSS, the

Table 4.6. The characteristics of designed level-shifters. The power numbers are evaluated by applying a clock signal of 1 MHz with a load capacitance of 5 fF operating at the corresponding stack voltage levels in the typical corner, 25°C.

Level-shifter	Average delay	Area per cell	Leakage power	Total power
TOP-MID	6.0 ns	$8.9216\mu m^2$	50 pW	8 nW
MID-TOP	4.0 ns	$8.2688\mu m^2$	150 pW	15 nW
MID-BOT	1.0 ns	$1.088\mu m^2$	800 pW	12 nW
BOT-MID	0.7 ns	$1.088\mu m^2$	800 pW	16 nW

two level-shifters can be used with the same ground reference. Hence, the designed chip can work in both voltage stacked mode and flat mode. Actually, the two level-shifters behave as buffer cells if the two power domains are connected in flat mode. The level-shifters were characterized and included in the physical design flow. The total number of level-shifters used between the TOP and MID stacks is 248, between the BOT stack and IO-pads is 18, whereas between MID and BOT stacks is 2080, out of a total of 390k gates in the design. In low voltage design, the level-shifters between TOP and MID domains are required in any case. The level-shifters between MID and BOT domains bring area overhead of ∼0.18%. The worst case timing overhead is 1.6% for the level-shifter between TOP and MID stacks.

4.4 Ultra-low voltage physical implementation

A 28-nm FDSOI 8-track standard cell LVT library was used for the implementation. The LVT standard cells show better process variation tolerance as compared to RVT standard cells in the near/sub-V_{th} region. The PB16 poly biasing standard cells were not used due to their relatively high threshold voltage. To implement the low voltage logic stacks, a pruned standard-cell library was generated by selecting cells from the foundry-provided 0.9 V libraries, which can operate reliably and efficiently at 0.4 V across all PVT corners (as partially discussed in Chapter 3: Library filtering). The generated standard cell library includes cells with transistor stacks of up to 2 transistors to keep the noise margin high while guaranteeing a high yield in the ultra-low voltage. In the pruned standard cell library, a limited set of driving strengths is allowed to keep the instantaneous switching power under control. Basically, low-drive strength cells feature a low switching current, resulting in better stability of the voltage stacks. The designed level-shifters were also characterized and included in the final library. The final library consists of ∼800 standard cells that were characterized

at multiple PVT corners using Cadence Liberate. The libraries were characterized for no FBB and with 200 mV FBB for the NMOS transistors.

The PVT corners used for the physical design are at 0.4 V±10%, SS, TT, FF corners, at 0°C, 25°C, 80°C temperatures, and min/max resistance/capacitance configurations. The PVT corners used for the foundry SRAMs were 0.9 V±10%, SS, TT, FF corners, at -40°C, 25°C, 120°C temperatures. The full SoC was implemented using Cadence Genus compiler for physical synthesis and Cadence Innovus for placement and routing with the common power format (CPF)-based multi-mode multi-corner (MMMC) approach. For the MMMC timing closure, a slow PVT corner (SS, 0.4 V, 0°C) was considered for the low voltage stacks. The hold violation fixing was performed for multiple PVT corners from SS, 0.4 V, 0°C up to TT, 0.6 V, 25°C. The PB16 delay cells from the clock library were used to fix hold violations. During logic synthesis, automatic clock gating was enabled and the synthesis tool has automatically inserted clock gating cells in the design. Automatic clock gating is a technique for power reduction in which unnecessary activity at the clock pin of flip-flops is avoided by gating the clock at some flip-flops based on the data path. The total number of gates in the design is 350k. The design is composed of 91% of PB10, 6% of PB16, 2% of PB4, and 1% of PB0 logic gates.

The BrainWave platform is split into five voltage domains. A voltage domain for IO-cells, a default core voltage domain for interfacing with IO-cells (0.8 V), the TOP voltage domain for SRAM (0.9 V), and the ultra-low voltage MID/BOT domains for the logic circuits operating at 0.4 V. The standard design flow using a CPF file to structure the power domains and low-power cells in the design. The CPF file enables the design flow to insert appropriate level-shifters between the defined voltage domains. A dense power grid is laid out to minimize the IR drop across the entire floor-plan for less than 10% drop.

The clock tree synthesis for near/sub-V_{th} operation across the stacked power domains is important. The utilization of the lowest threshold voltage is recommended for clock tree synthesis. Only PB0 and PB4 poly-biased selected high strength inverters, buffers, and clock-gating cells from the clock library were used. The balanced transition times of these cells reduce the clock tree variability. The relatively lower threshold voltage of PB0 and PB4 cells as compared to PB10 and PB16 cells improves the insertion delay and clock skew at low voltage. Cells with higher drive strength were allowed for the clock tree synthesis to further improve the insertion delay. In this voltage stacked system, each power domain has its local clock tree. The clock signal in our voltage stacked power domains requires an appropriate voltage level. As shown in Fig. 4.15, the MID and BOT stacks in the design do not require clock signal level shifting. However, one level-shifter is required for the clock signal to the TOP stack. The clock tree is local within a power domain while the skew is minimized

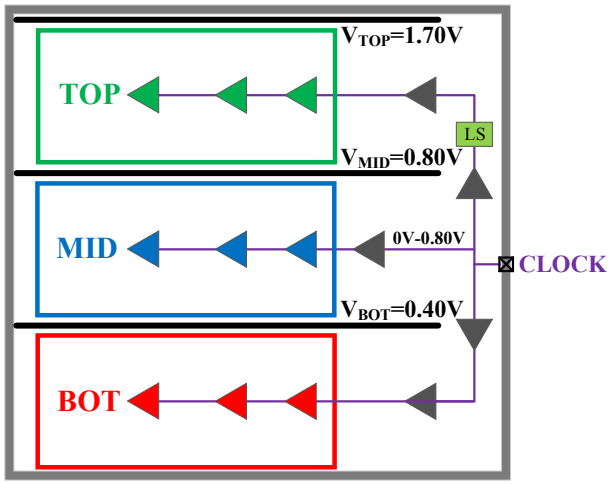


Figure 4.15. Clock tree for our three level voltage stacked design.

over the entire chip. In [142], the clock tree requires multiple level-shifters as the clock signal is crossing from one power domain to the other multiple times. In our case, this is not necessary. The clock tree insertion delay and clock skew across the voltage stacks in the chip are within 3% and 2% of the clock period, respectively.

The chip layout and floor-planning are shown in Fig. 5.1. The placement of the bond-pads, IO-pads, two CS, the four ABB-based voltage controllers across the chip, and the voltage stack partitioning are depicted in the chip layout. The chip size is 1.18 mm×1.16 mm. For the low voltage input/measurement, the analog type of IO-cells is used from the IO library. All the IO-cells are internally protected from electrostatic discharge. The power pins of each power domain are routed outside the chip. Hence, the chip can be configured to operate in either voltage stacked or flat mode, thanks to the designed level-shifters for voltage stacking. In the flat mode, the separate power supplies for the SRAMs and logic circuits are provided externally.

4.5 Summary

A new approach for implementing digital ICs operating in the near/sub-V_{th} region is presented in this chapter. A three-level voltage-stacked system consisting of a RISC-V core and a CGRA accelerator is implemented in a 28-nm FDSOI technology. The near/sub-V_{th} region operation of the stacks results in the balancing of the stacks with simpler circuits. The presented voltage-stacking technique operates using a

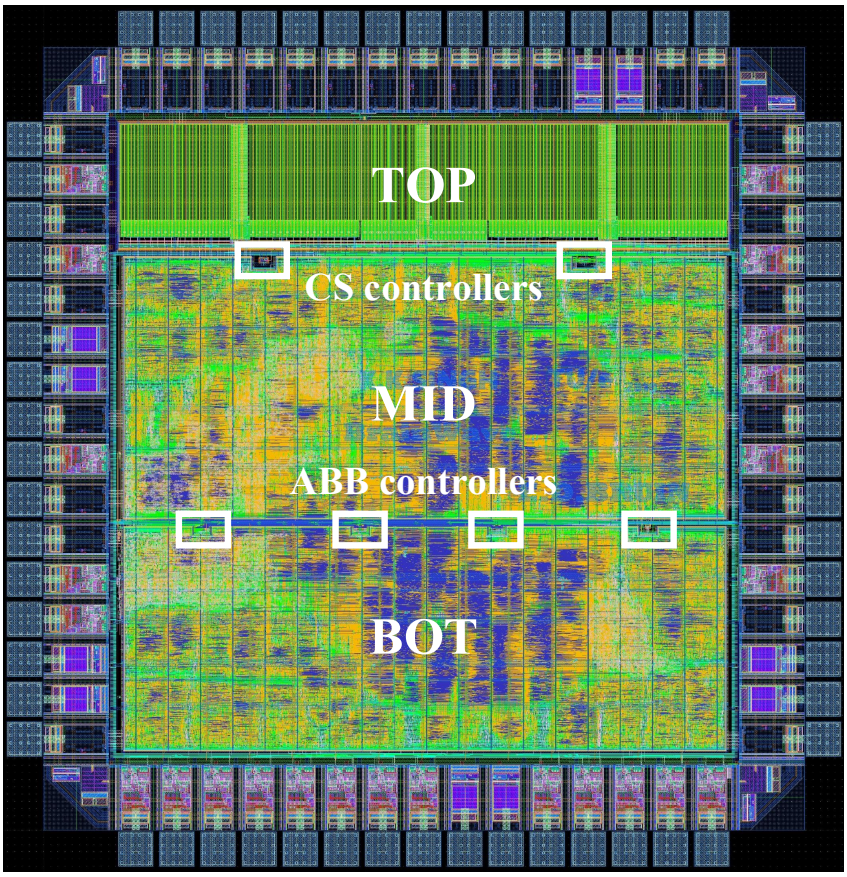


Figure 4.16. Chip layout and floor-planning showing the stacks and placement of voltage controllers.

single supply voltage of 1.7 V for the core as well as the IO subsystem, eliminating the need for additional voltage regulators.

The stack balancing technique uses a current sink controller and an ABB controller. The ABB voltage controller balances the MID and BOT stacks by changing the body-bias voltage to balance the current. These voltage controllers are able to balance the intermediate node voltage variation within a tolerance limit of ±5% of the voltage. The "converter-free" implementation results in significant area savings.

Chapter 5

Silicon Measurement

In this chapter, we will present the silicon measurement of the proposed voltage-stacked system in Chapter 4. The chip micrograph is shown in Fig. 5.1. The chip is wire bonded to a 68-pin J-Leaded Ceramic Chip Carrier (JLCC) package to interface with a simple test set-up. The power and body-bias pins of the three voltage domains are separately routed outside the chip on the left and right side of the chip for monitoring as well as for external stimulus. The 17-GPIO pins are routed outside the chip mostly from the bottom side of the chip. The SPI/QSPI/UART/I2C/JTAG peripherals are multiplexed with the GPIOs. The designed chip does not contain circuits for clock generation and power management. Therefore, the test set-up provides connectors and test points for external sources and measurements. The designed test printed circuit board (PCB) supports the switching from the voltage-stacked mode to flat-mode operation using jumpers settings. The chip is programmed using the JTAG interface. An Altera FPGA is used for programming the chip and interfacing for the test data. The test-board consists of bi-directional level-shifters for interfacing with the FPGA-board, since the FPGA GPIO's operate at 3.3 V and the chip operates at 1.7 V. The test set-up is shown in Fig. 5.2.

The proposed voltage-stacked system is compared using various measured parameters. The key measurement parameter to evaluate the voltage-stacked system is the total system efficiency ($\eta_{sys-stack}$), which is defined by

$$\begin{aligned}
\eta_{sys-stack} &= \frac{P_{mem} + P_{logic}}{P_{total-stack}}, \\
P_{mem} &= (V_{TOP} - V_{MID}) \times I_{TOP}, \\
P_{logic} &= (V_{MID} - V_{BOT}) \times I_{MID} + V_{BOT} \times I_{MID}, \\
P_{total} &= V_{TOP} \times I_{in}, \\
\eta_{sys-stack} &= \frac{(V_{TOP} - V_{MID}) \times I_{TOP} + V_{MID} \times I_{MID}}{V_{TOP} \times I_{in}},
\end{aligned} \tag{5.1}$$

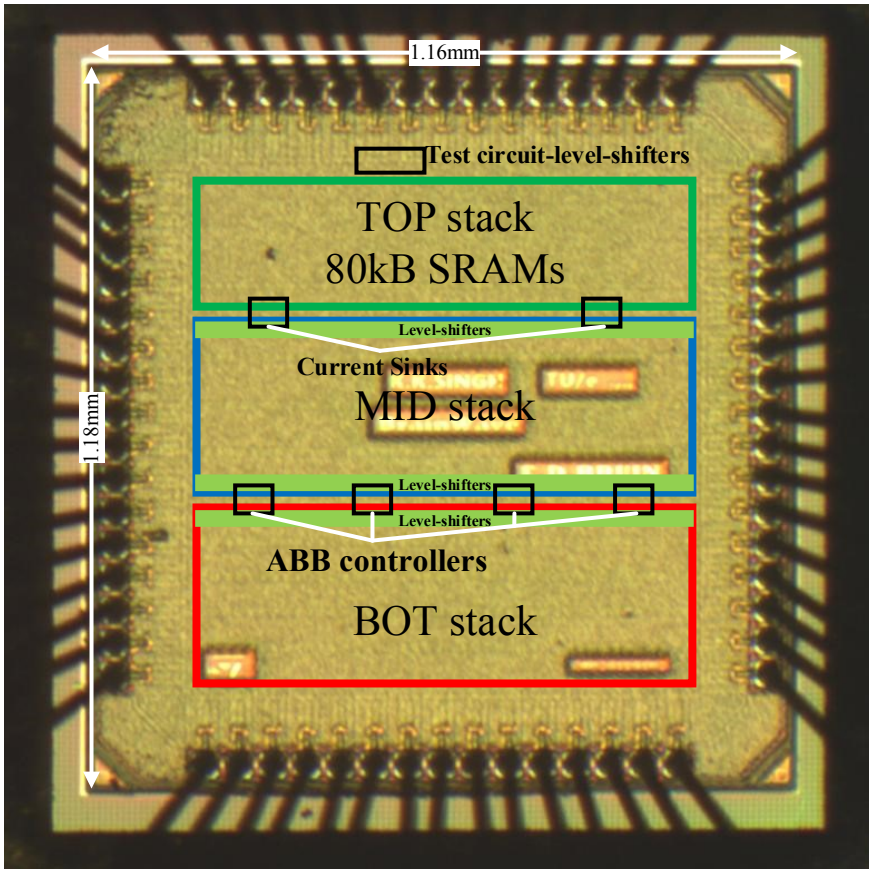


Figure 5.1. The wire-bonded chip micrograph showing the stacks and placement of voltage controllers.

where I_{in} is the current consumption from the supply voltage V_{in} (=V_{TOP}). The TOP stack current consumption is I_{TOP} (= $I_{in} - I_{cs}$). I_{cs} is the current consumption of the CS voltage controllers. I_{MID} is the MID stack current consumption, which is the same as the I_{BOT} BOT stack current consumption. The system efficiency is the maximum when I_{TOP} is equal to I_{MID}. The system efficiency equation includes the power consumption overhead of the on-chip current sink controllers and the ABB based voltage controllers. The power consumption overhead of the level-shifters and external voltage reference supplies are ignored for simplicity. All the required cir-

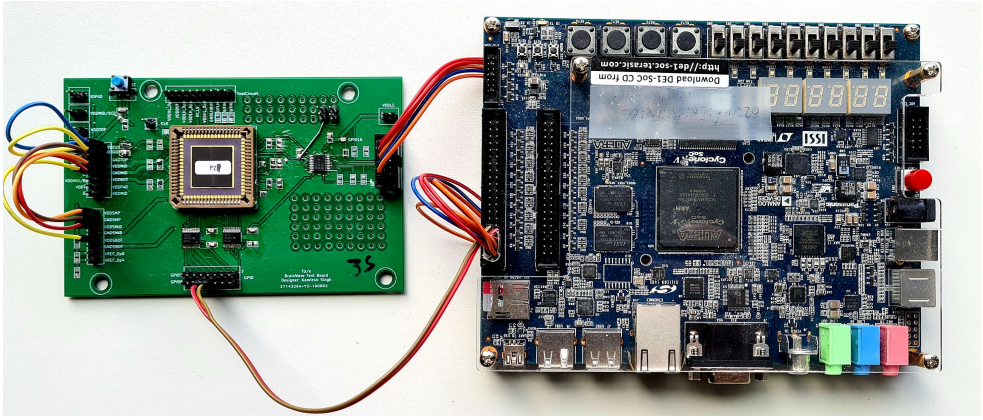


Figure 5.2. Measurement set-up for chip performance evaluation.

cuits are implemented inside the chip except the voltage reference for the voltage controllers.

For the flat-mode shown in Fig. 1.5 (Chapter 1), the system efficiency is given by

$$\begin{aligned} \eta_{sys-flat} &= \frac{P_{mem} + P_{logic}}{P_{total-flat}}, \\ P_{total-flat} &= \frac{P_{mem} + \frac{P_{logic}}{\eta_{LDO}}}{\eta_{SCVR}}, \\ \eta_{sys-flat} &= \frac{\eta_{SCVR}(P_{mem} + P_{logic})}{P_{mem} + \frac{P_{logic}}{\eta_{LDO}}}, \end{aligned} \tag{5.2}$$

where η_{SCVR} is the conversion efficiency of the SCVR and η_{LDO} ($\propto \frac{V_{logic}}{V_{mem}}$) is the conversion efficiency of the LDO. The flat-mode system efficiency is limited by the SCVR conversion efficiency in the worst case. The flat-mode system efficiency is much lower than η_{SCVR} due to significant conversion loss in LDO. In this thesis, the assumed efficiency of the SCVR is 85% [89] for converting 1.7 V to 0.9 V and for the LDO is 44% (0.4/0.9) for converting 0.9 V to 0.4 V. Additionally, the system energy consumption and area overheads are compared in the voltage-stacked mode and flat-mode.

In this chapter, firstly, the silicon characterization of the designed voltage-stack intermediate rail balancing controllers is presented. The CS and ABB controllers are characterized by forcing external current/test stimulus. Subsequently, the silicon measurements of the chip by executing multiple signal processing benchmarks are presented. Furthermore, the voltage-stacking technique is compared against the con-

ventional flat-mode implementation. The measurement results are compared against the state-of-the-art ultra-low-power systems. Finally, the measurement results of multiple dies for voltage and temperature variations are presented.

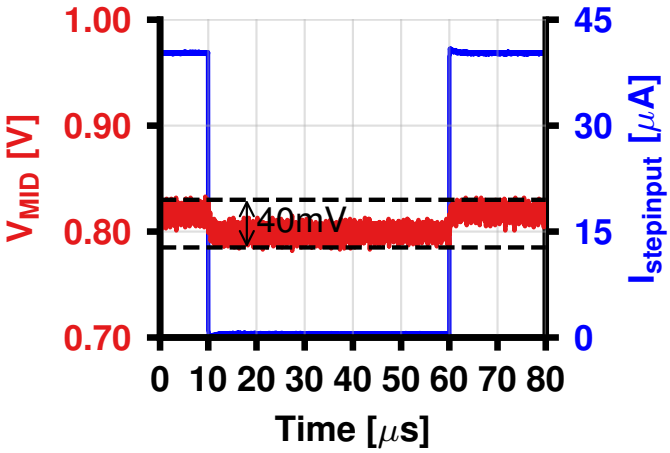
5.1 Voltage-stacking measurement

5.1.1 Voltage-stack balancing controller measurement

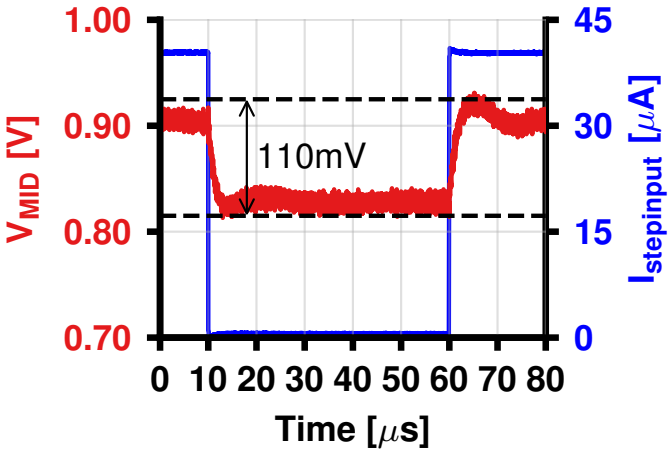
CS controller measurement results: The CS controller was designed to control the voltage swing of the V_{MID} rail to within ±5% of 0.8 V. When a step current change ΔI_{TOP} occurs, the voltage at V_{MID} changes by ΔV_{MID} in a time interval T_r (= $C_{decap}\Delta V_{MID}/\Delta I$) for a decoupling capacitance C_{decap} [111]. T_r, consequently, determines the response time of the CS controller to achieve the voltage drop of ΔV_{MID}. It is measured by forcing a step current of 40 μA externally into V_{MID} as was shown in Fig. 4.6. The forced step current and the measured V_{MID} rail voltage are shown in Fig. 5.3a. The measured response time and settling time for rise and fall of the step current are <500 ns and <2 μs, respectively. The peak to peak voltage variation of the V_{MID} rail is 40 mV which is within ±5% of 0.8 V. The V_{MID} balancing without the CS controller is shown in Fig. 5.3b. In this case, the measured peak to peak voltage variation is 110 mV. The total current consumption of the two CS controllers (I_{cs}) is 200 nA at 1.7 V worst case supply voltage.

ABB voltage controller measurement results: The ABB voltage controller controls the V_{BOT} rail between the MID and BOT stacks at 0.4 V. Similar to the CS controller, the response time (T_r) of the ABB controller to meet a voltage drop of ±5% depends on the current imbalance and the decoupling capacitor (= $C_{decap}\Delta V_{BOT}/\Delta I$). The characterization is done by forcing a step current externally with a peak to peak amplitude of 20 μA to unbalance V_{BOT}, as shown in Fig. 5.4. The measured response time is less than 1 μs. The settling time of V_{BOT} at 0.4 V is ∼10 μs. The peak to peak variation of V_{BOT} is ∼40 mV. Because of the use of body biasing, the controller keeps the voltage swing of V_{BOT} to within ±5% of 0.4V with a maximum performance loss of 5%/100 mV change in the NMOS body bias voltage. The performance loss occurs only when the body bias voltage goes below the initial balanced condition. Under these strained conditions, V_{BB_MID} drops by ∼130 mV, corresponding to a performance loss of ∼6%, see Fig. 5.4. The total current consumption for the four ABB controllers is ∼4 μA at 0.8 V. The current for the ABB controllers is provided by the TOP stack. Additionally, off-chip decoupling capacitors can be used to relax the design specification of controllers to meet the voltage drop requirement of the V_{BOT} and V_{MID} rails.

The performance of the ABB controller is also evaluated by stressing the power



(a)



(b)

Figure 5.3. Silicon measurement of the CS controllers. (a) The measured voltage regulation of V_{MID} rail by the CS controller upon forcing a step current externally and (b) V_{MID} rail regulation without CS controller. The intermediate voltage rail droops are obtained without any external decoupling capacitor.

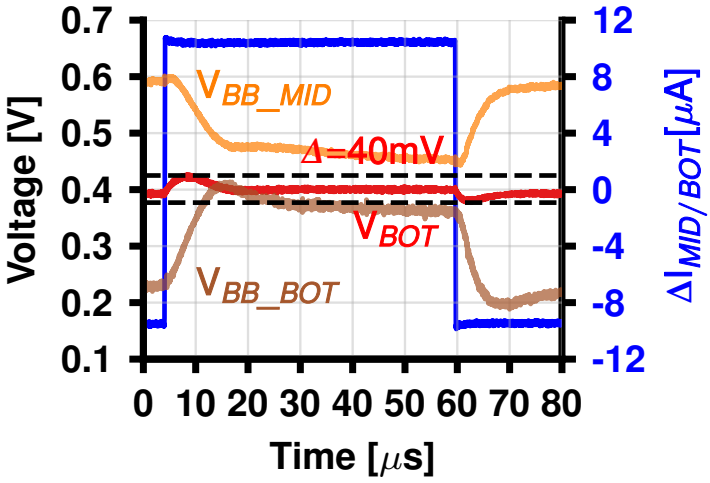
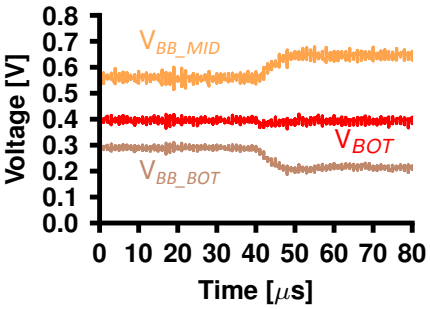
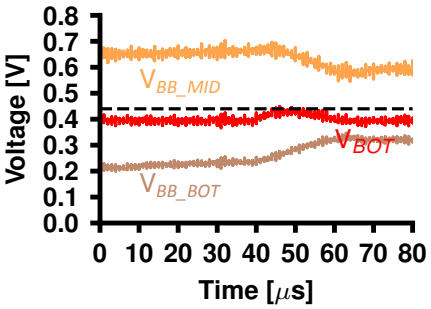


Figure 5.4. The silicon measurement of the voltage regulation of V_{BOT} rail by the ABB controller upon forcing a step current externally. The intermediate voltage rail droops are obtained without any external decoupling capacitor.

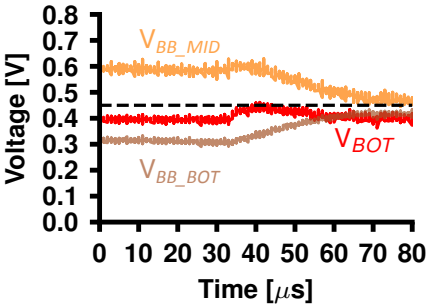
domains in multiple modes instead of forcing current externally. The multipliers in each of the scalar cores in the MID and BOT stacks are used to run multiplication of random data to create different workload modes. The different possible modes are IDLE mode (no activity except clock), MID-active (two multipliers in the MID stack scalar cores are active), and BOT-active (two multipliers in the BOT stack are active). The designed voltage stack is very stable while continuously operating in one mode. However, switching from one mode to another mode create stress in the voltage stack. The voltage stack behavior in different stress conditions is shown in Fig. 5.5. The V_{BOT} rail while continuously operating in a mode remains stable at 0.4 V. The voltage on the V_{BOT} rail changes while switching between different modes and then stabilizes at 0.4 V. In most of the mode switching the V_{BOT} voltage varies (ΔV_{BOT}) within 10 mV, with a settling time of $<2\,\mu s$. The transition from BOT-active to IDLE mode and IDLE to MID-active mode show relatively higher variation of 30 mV, 40 mV, respectively. In this case, the V_{BOT} settling time to 0.4 V is within $\sim 20\,\mu s$. The worst case is when the workload changes from MID-active to BOT-active showing a ΔV_{BOT} of 80 mV and settling time of $\sim 40\,\mu s$. The switching from IDLE to BOTH-active mode causes $\sim$10 mV change on the V_{BOT} rail. The worst case load switching should be avoided while using the proposed voltage stacked system.



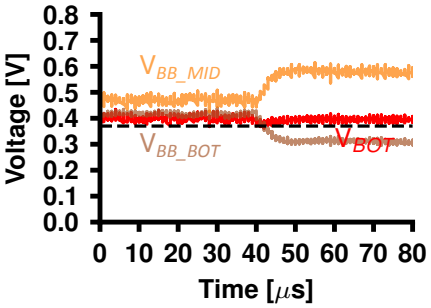
(a) IDLE to BOT-active mode ($\Delta V_{BOT} \approx 10\,\text{mV}$).



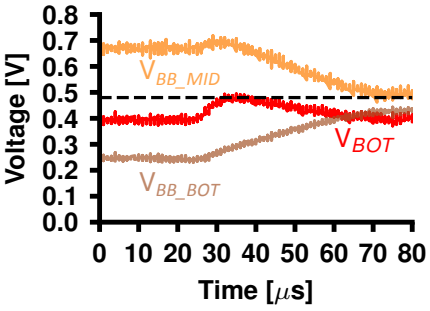
(b) BOT-active to IDLE mode ($\Delta V_{BOT} \approx 30\,\text{mV}$).



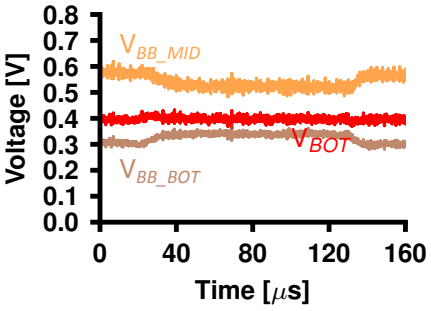
(c) IDLE to MID-active mode ($\Delta V_{BOT} \approx 40\,\text{mV}$).



(d) MID-active to IDLE mode ($\Delta V_{BOT} \approx 10\,\text{mV}$).



(e) MID-active to BOT-active mode ($\Delta V_{BOT} \approx 80\,\text{mV}$).



(f) IDLE to BOTH-active to IDLE mode ($\Delta V_{BOT} \approx 10\,\text{mV}$).

Figure 5.5. Silicon measurement of the ABB controllers in different scenario. The intermediate voltage rail droops are obtained without any external decoupling capacitor.

Application-level voltage smoothing techniques can be used to also mitigate the stack balancing issue [140].

5.1.2 Chip performance evaluation for various benchmarks

To demonstrate the dynamic behavior of the voltage stacked system, a 16-bit fixed-point 32×32 *MatMul*, a 16-bit 10^{th} order (21 taps) Butterworth filter, and a data-intensive merge-sort application were used for benchmarking. The applications stresses the stacks by varying the workload among four different modes of operation as follows:

- IDLE mode: no activity except for the clock.
- RISC-V active: RISC-V core active CGRA idle (the MID stack needs more current than the BOT stack).
- CGRA active: RISC-V core sleep CGRA active (MID and BOT stacks active, approximately equal distribution of workload between the MID and BOT stacks).
- BOTH active: both RISC-V core and CGRA are active (the MID stack needs more current than the BOT stack).

For the sake of measurement, the applications on the CGRA run multiple times to have equal runtime as on the RISC-V core.

Voltage stack balancing: The distinct voltage rails while running *MatMul* at 2 MHz are shown in Fig. 5.6. The measurement shows balanced voltage stacks with V_{MID} balanced at an average voltage of 0.79 V and V_{BOT} balanced at an average voltage of 0.39V. The measured average variation of V_{MID} and V_{BOT} are 20 mV and 10 mV, respectively, across varying workloads.

The current consumption of the TOP and MID stacks is summarized in Fig. 5.7. In the RISC-V active mode, the instruction/data SRAM memories in the TOP stack are accessed frequently. Hence, the TOP stack current consumption I_{TOP} is higher than the current consumption of the logic stacks (I_{MID}). In the CGRA-active mode, the RISC-V core is in sleep mode with no activity to the instruction memory. The CGRA uses the local data/instruction standard cell based memories for processing, resulting in relatively less power consumption of instruction and data SRAMs in the TOP stack. Therefore, the difference of the current consumption between the TOP and MID stacks is low. In the BOTH active mode, the SRAMs in the TOP stack, as well as the RISC-V core and CGRA in the MID and BOT stacks, are active.

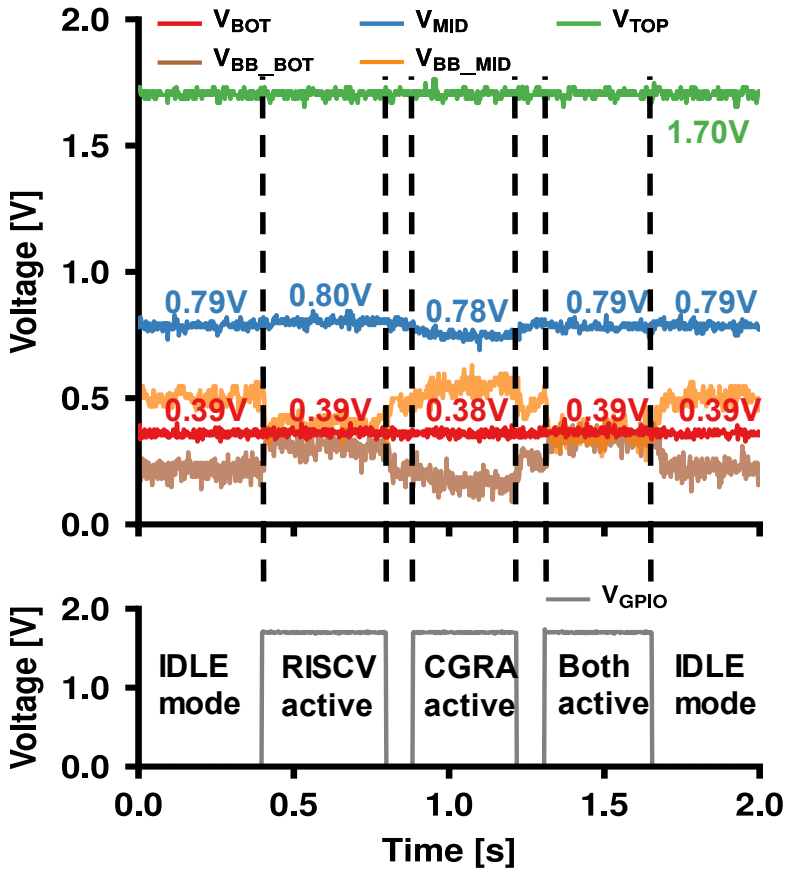
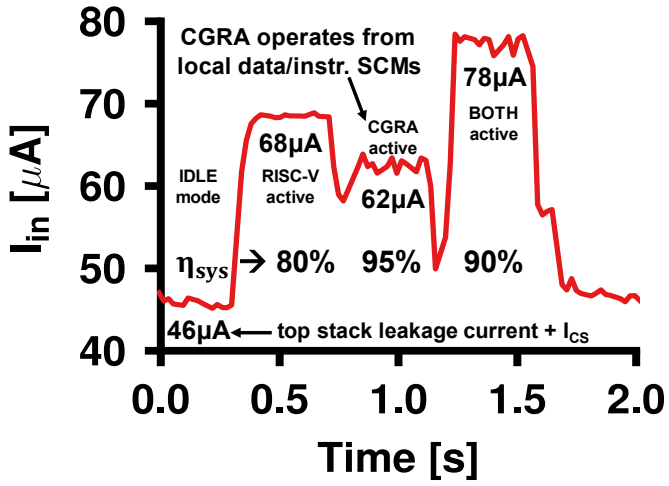
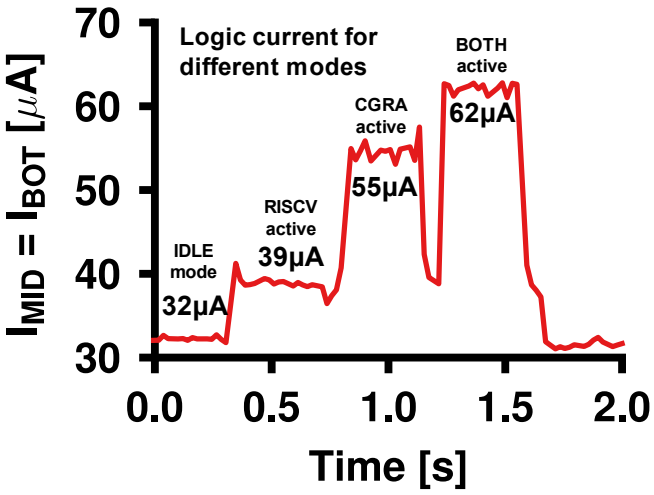


Figure 5.6. Measurement of balanced stack operating at 2 MHz while running *MatMul*.

CS controller operation and system efficiency: As shown in Fig. 5.7, in the RISC-V-active mode, the current consumption of the TOP stack is 68 μA, whereas the current consumption of logic stacks is 39 μA. Therefore, a higher amount of current from the TOP stack is sunk by the CS controllers to balance the V_{MID} rail at 0.8 V. This results in a relatively low total system efficiency. In the CGRA-active mode, the power consumption of the logic stacks is higher than in the RISC-V-active mode. Hence, most of the current from the TOP stack is utilized by the logic stacks. Therefore, the system efficiency is higher in the CGRA-active mode when compared



(a) I_{TOP}



(b) I_{MID}=I_{BOT}

Figure 5.7. The measured current flowing through TOP and MID/BOT stacks with varying workload.

to the RISC-V-active mode. The current consumption of the ABB controllers is provided by the TOP stack. Therefore, a part of the current difference between the TOP

Table 5.1. The ABB controller outputs V_{BB_BOT} and V_{BB_MID} are shown regulating V_{BOT} for different workloads.

	IDLE mode	**RISC-V active**	**CGRA active**	**BOTH active**
	Initial	$I_{MID}>I_{BOT}$	$I_{MID}<I_{BOT}$	$I_{MID}>I_{BOT}$
$\mathbf{V_{BB\_MID}}$	500 mV	410 mV (Δ =-90 mV)	530 mV (Δ =30 mV)	390 mV (Δ =-110 mV)
$\mathbf{V_{BB\_BOT}}$	240 mV	370 mV (Δ =130 mV)	200 mV (Δ =-40 mV)	380 mV (Δ =140 mV)

and MID stacks is used by the ABB controllers. The achieved peak system efficiency is limited by the current consumption of the ABB controllers, used for balancing the V_{BOT} rail at 0.4 V. In the BOTH active mode, the TOP stack consumes more current as compared to other modes. However, most of the current from the TOP stack is consumed by the active RISC-V core and the CGRA in the MID and BOT stacks, resulting in medium system efficiency.

ABB controller operation and system efficiency: Recall that the body bias voltage of the MID stack NMOS transistors (V_{BB_MID}) and the body bias voltage of the BOT stack NMOS transistors (V_{BB_BOT}) change in opposite directions with the varying workload to balance the V_{BOT} rail at 0.4 V, see Fig. 5.6. The operation of the ABB controller is summarized in Table 5.1. In IDLE mode, V_{BB_MID} and V_{BB_BOT} stabilize at 0.5 V and 0.24 V, respectively. The initial condition shows that the NMOS transistors in the MID stack are forward biased by 100 mV, and the NMOS transistors in the BOT stack are forward biased by 240 mV. This balances the initial imbalance between the MID and BOT stacks due to process variability. Subsequently, when the workload varies, the body bias voltage adapts to keep the stacks balanced. In the RISC-V active mode, the RISC-V core in the MID stack is active, increasing the current consumption of the MID stack. The increased current in the MID stack forces the BOT stack to consume the same current. Hence, the V_{BB_BOT} shows increased forward bias on the BOT stack and also reduced forward bias on the MID stack to balance the V_{BOT} rail at 0.4 V. As shown in Fig. 5.6, the V_{BB_BOT} voltage increases to 370 mV from 240 mV, whereas the V_{BB_MID} voltage decreases to 410 mV from 500 mV. The behavior of the ABB controller in BOTH active mode is similar to the RISC-V active mode. In CGRA-active mode the MID and BOT stacks are balanced and hence the voltage of the V_{BB_MID} and V_{BB_BOT} are similar as of the IDLE mode as shown in Table 5.1.

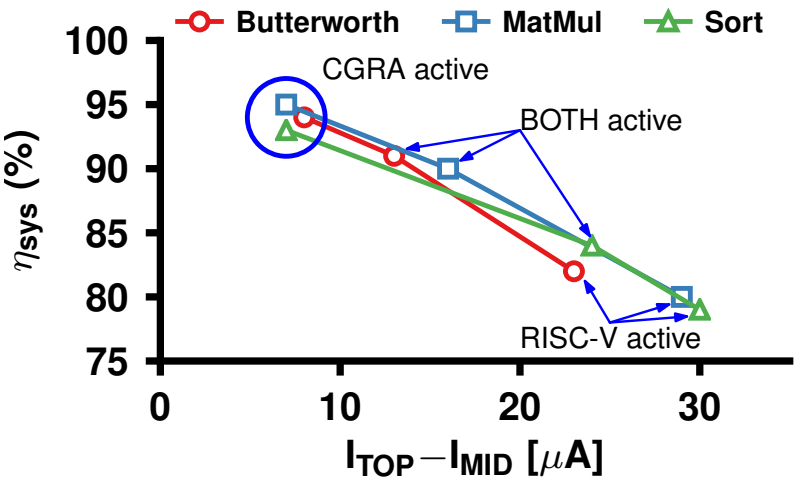


Figure 5.8. The measured system efficiency vs current difference between the TOP and MID stacks.

System efficiency measurement: The measured system efficiency for multiple benchmarks is shown in Table 5.2. The system efficiency versus the current difference between the TOP and MID stacks is shown in Fig. 5.8. The highest efficiency is in

Table 5.2. Benchmarks with V_{TOP}=1.70 V and V_{MID}=0.80 V. The system efficiency is calculated using equation (5.1).

Configuration	Benchmarks	I_{in}	I_{MID}	η_{sys}
RISC-V active	MatMul	68 μA	39 μA	80%
	Butterworth	61 μA	38 μA	82%
	Merge sort	68 μA	38 μA	79%
CGRA-active	MatMul	62 μA	55 μA	**95%**
	Butterworth	58 μA	50 μA	**94%**
	Merge sort	50 μA	43 μA	**93%**
Both-active	MatMul	78 μA	62 μA	90%
	Butterworth	69 μA	56 μA	91%
	Merge sort	72 μA	48 μA	84%
Idle mode	MatMul	46 μA	32 μA	86%
	Butterworth	45 μA	32 μA	86%
	Merge sort	45 μA	30 μA	84%

the CGRA-active mode. The measured system efficiency is greater than 93% with a maximum of 95%. The system efficiency is greater than 90% in the BOTH active mode except for sorting. The sorting algorithm running on the CGRA consumes relatively less power due to a large number of comparisons instead of multiplications.

5.1.3 Microcontroller configuration for specific applications

In this section, the energy efficiency of the system in voltage stacked mode is measured and also compared against the flat-mode. Without loss of generality, the energy efficiency is measured by running the *MatMul*, Butterworth filtering, and merge sort applications. Generally, the applications executing on the CGRA are ∼6–15× more energy-efficient than when running on the RISC-V core, as shown in Table 5.3. This is partly attributed to the number of clock cycles needed to run the application. For instance, the *MatMul* application running on the CGRA requires 24k cycles, whereas running on the RISC-V core requires 134k cycles. The system achieves a maximum frequency of 2.8 MHz at 1.7 V during the CGRA-active mode. The measured energy efficiency for the *MatMul* application is 35.1 MMAC/s/mW (million multiply-accumulation per second per milliWatt). The measured peak performance is 4 MMAC/s. Likewise, the energy efficiency of the sorting algorithm is 4.6 MCMP/s/mW (million comparisons per second per milliWatt). In our platform, the RISC-V can only execute 1 operation/cycle, whereas the CGRA can perform a maximum of 20 operations/cycle. For the 32×32 matrix multiplication application, the CGRA requires 24k cycles whereas the RISC-V requires 134k cycles. In voltage stacked mode, the energy consumption when the CGRA is active is 39.2pJ/cycle, and when the RISC-V is active it is 47.6pJ/cycle. Therefore, the total energy consumed for the matrix multiplication by the CGRA (39.3pJ/cycle×24k) is 6.8× lower than for the RISC-V (47.6pJ/cycle×134k). Hence, the recommendation is using CGRA for computation instead of RISC-V.

Voltage stacked vs flat-mode implementations: The chip can be externally configured to operate in flat-mode by connecting the power domains (stacks) in parallel. The conversion to the flat-mode is possible due to the dual behavior of the designed level-shifters. The flat-mode behavior is measured by connecting an external supply of 0.9 V to the TOP stack, 0.4 V to logic stacks, 1.7 V for IO-pads, and forward biasing the NMOS transistors by 200 mV, in order to match the default voltage stacked condition. The measurement results show that the chip in the flat-mode achieves a higher frequency as compared to the voltage stacked mode due to constant forward body biasing. The measured energy efficiency for the benchmarks with and without accounting for voltage conversion losses are tabulated in Table 5.3. The assumed con-

version loss for the flat-mode is as shown in Fig. 1.5, where the SCVR has a conversion efficiency of 85% [89] and the LDO has a conversion efficiency of 44% ($\propto V_{out}/V_{in}$). The energy efficiency in the voltage stacked mode is up to 1.7 × higher as compared to the flat-mode. The energy consumption per cycle for the voltage stacked mode is up to 42% lower as compared to the flat-mode. Moreover, the energy consumption reduces by 42% for *MatMul* application in the CGRA-active mode. Furthermore, the measurement of 10 chips from two different lots provided by the foundry shows an average energy consumption of 38.8 pJ/cycle while executing *MatMul* in the CGRA-active mode. The energy efficiency in the voltage-stacked mode is on average 1.6 × higher and the energy consumption is on average 37% lower as compared to the flat-mode implementation. We also measured the chip in the flat-mode without FBB, the average energy consumption is 59.0 pJ/cycle which is ∼4% lower as compared to the flat-mode with a 200 mV FBB. The result of chip measurement without FBB is shown in Table 5.4

5.2 Voltage scaling measurement

The proposed voltage-stacked system allows voltage scaling within a limited range. The voltage scaling range is limited by the design of controllers to balance the intermediate rails and the assumed current consumption ratios between stacks while partitioning. The voltage-stacked system is measured by varying the supply voltage of the stack from 1.8 V down to 1.5 V in the steps of 100 mV by executing the *MatMul* kernel. Additionally, to balance the voltage-stack the external reference voltages are adjusted to achieve suitable operating intermediate rail voltages. The measured energy consumptions, operating frequencies, and voltage across each stack in the system are shown in Fig. 5.9. The measurement results show that when the voltage is scaled the voltage across the BOT stack has to be adjusted to $\sim V_{MID}/2$ to keep the stack balanced. Therefore, the voltage across the MID and BOT stacks scale from ∼450 mV down to ∼350 mV when the supply voltage is scaled from 1.8 V down to 1.5 V, respectively.

In the flat-mode, the supply voltage is possible to be scaled from 0.9 V down to 0.35 V, thanks to the design of the level-shifters. Additionally, as per the data-sheet of SRAMs, the safe supply voltage of operation is up to 0.8 V. Therefore, to have a fair comparison, the supply voltage of SRAM is reduced to 0.8 V in the flat-mode of operation. The measured energy consumption and frequency of operation are shown in Fig. 5.10.

In the voltage-stacked mode, the measured minimum energy consumption is 38 pJ/cycle at 1.7 V, while the lowest power consumption is 55.6 μW at 1.5 V. The maximum operating frequency achieved is 3.5 MHz at 1.8 V, which is limited by the

Table 5.3. Measurement of energy efficiency with the voltage stack operating at V_{TOP}=1.7 V, and the flat-mode operating at V_{mem}=0.9 V and V_{logic}=0.4V with fixed 200 mV FBB. In the flat-mode, the conversion losses are the same as depicted in Fig. 1.5. The assumed efficiency of the SCVR is 85% [89] for converting 1.7 V to 0.9 V and for the LDO is 44% for converting 0.9 V to 0.4 V. The shown energy efficiency number in parentheses is for the flat-mode without considering conversion losses.

Mode	**Benchmarks**	**Voltage stacked mode**					**Flat-mode**				
		Freq. (MHz)	**I_{in} (μA)**	**I_{MID} (μA)**	**Energy (pJ/cycle)**	**Energy efficiency**	**Freq. (MHz)**	**I_{mem} (μA)**	**I_{logic} (μA)**	**Energy (pJ/cycle)**	**Energy efficiency**
CGRA active	MatMul	2.8	65	57	39.2	**35.1** MMAC/s/mW	3.0	62	128	67.5 (35.7)	**20.5** (38.8) MMAC/s/mW
	Butterworth	2.8	70	62	42.5	**31.7** MMAC/s/mW	3.0	67	127	68.9 (37.0)	**19.5** (36.4) MMAC/s/mW
	Merge sort	2.7	54	45	34.0	**4.6** MCMP/s/mW	4.1	61	125	48.4 (25.6)	**3.2** (6.1) MCMP/s/mW
RISC-V active	MatMul	3.0	84	48	47.6	**5.2** MMAC/s/mW	6.0	160	128	51.1 (32.5)	**4.8** (7.6) MMAC/s/mW
	Butterworth	3.0	76	47	43.1	**2.0** MMAC/s/mW	5.5	132	118	48.4 (30.2)	**1.8** (2.8) MMAC/s/mW
	Merge sort	3.0	86	48	48.7	**0.8** MCMP/s/mW	7.6	192	143	47.0 (30.3)	**0.8** (1.3) MCMP/s/mW

Table 5.4. Chip measurement in the flat-mode without 200mV FBB. In the flat-mode the conversion losses are the same as depicted in Fig. 1.5. The assumed efficiency of the SCVR is 85% [89] for converting 1.7 V to 0.9 V and for the LDO is 44% for converting 0.9 V to 0.4 V.

	Benchmarks	I_{mem} 0.9V (μA)	I_{logic} 0.4V (μA)	Frequency (MHz)	Energy (pJ/cycle)	Energy efficiency
CGRA active	MatMul	62	110	2.9	63.2	21.9 MMACs/mW
	Butterworth	66	109	2.7	69.1	19.5 MMACs/mW
	Sorting	61	103	3.9	44.8	3.5 MCMPs/mW

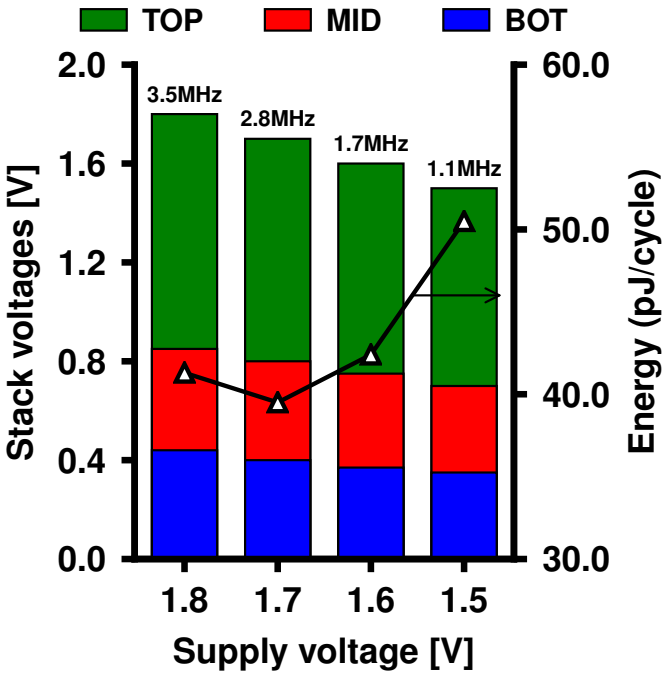


Figure 5.9. Voltage scaling of stacked system and energy variation.

current sourcing capacity from the SRAMs. In the flat-mode of operation, the measured minimum energy consumption is 43.4 pJ/cycle at 0.5 V for the logic circuit, which is 12% higher as compared to the voltage-stacked mode. It is interesting to

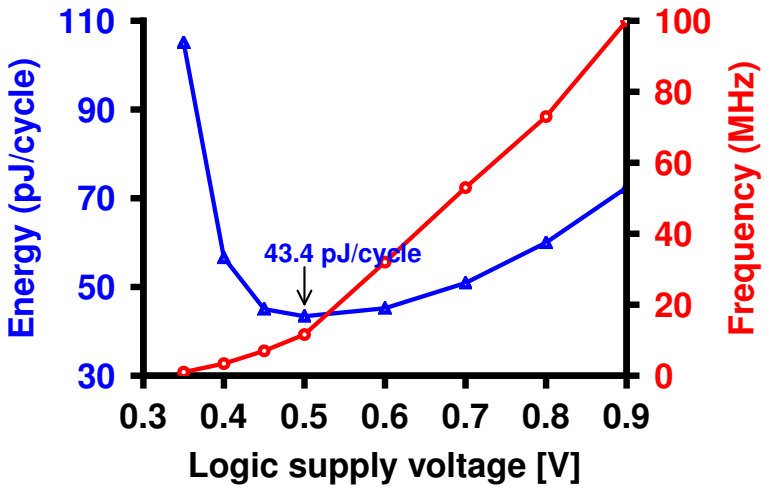


Figure 5.10. Measured energy consumption vs supply voltage in flat-mode. The supply voltage for the SRAM blocks is 0.8 V. For the logic circuit, the assumed conversion losses of the SCVR is 85% [89] for converting 1.7 V to 0.9 V and for the LDO is $V_{out}/0.9V$.

note that in the flat-mode of operation, the minimum power consumption is 105 μW at a logic supply of 0.35 V, which is 48% higher as compared to the voltage-stacked mode with a similar achieved operating frequency of ~1 MHz. Moreover, at the minimum power consumption operation, the energy consumption is 52% lower in the voltage-stacked mode as compared to the flat-mode of operation.

5.3 Multiple-die measurement

The designed systems are operating at ultra-low supply voltage, variations (inter-die and intra-die) can have a big impact on the performance of the chip. In this section, the silicon measurement results of 10-bonded ICs from two different lots are presented. The energy consumption and operating frequency for different supply voltages are measured for all the ICs. In the voltage-stacked and flat-mode mode, the measured energy consumption and the average operating frequency vs the logic supply voltage are shown in Fig. 5.11 and Fig. 5.12, respectively. The average minimum energy consumption in voltage-stacked and flat-mode are 36 pJ/cycle, and 42 pJ/cycle, respectively. The relative minimum energy point is varying by 35% in

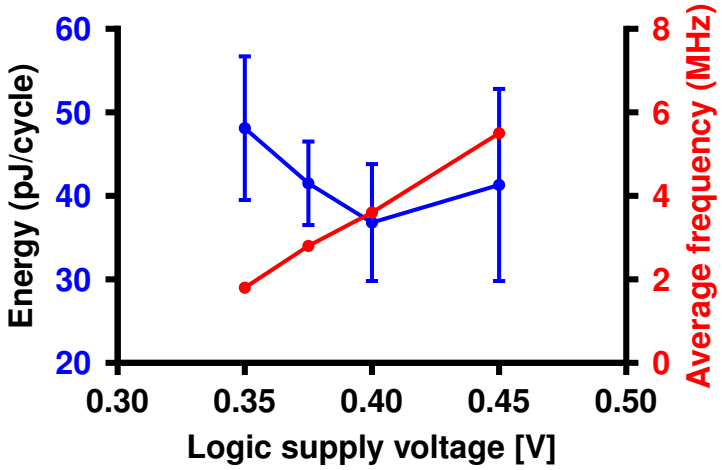


Figure 5.11. Multiple-die measurement from 2 different lots in the voltage-stacked mode.

the voltage-stacked mode and 41% in the flat-mode. Additionally, the relative variation of the energy consumption at the minimum power point is significantly lower for the voltage-stacked mode (∼33%) as compared to the flat-mode (40%) of operation. Therefore, the voltage-stacked system benefits by reducing the effects of process variations [112].

5.4 Temperature measurement

Apart from process variations, ambient temperature variations are an important source of system failure. The chip is measured to demonstrate the voltage-stack balancing for temperature variations. The proposed voltage-stacked chip is designed to operate over a temperature range of 0°C to 80°C. A Peltier element is used to change the temperature of the chip. The Peltier element can make a temperature difference between its two sides if enough current is applied. To set the temperature to a specific value, the applied current must be tuned accordingly. The temperature is sensed with a Thermocouple that is placed between the Peltier element and the chip. An Arduino platform reads the temperature and controls the current accordingly to reach the target temperature. For cooling the Peltier element below room temperature, the input terminals are reversed such that the current flows in the reverse

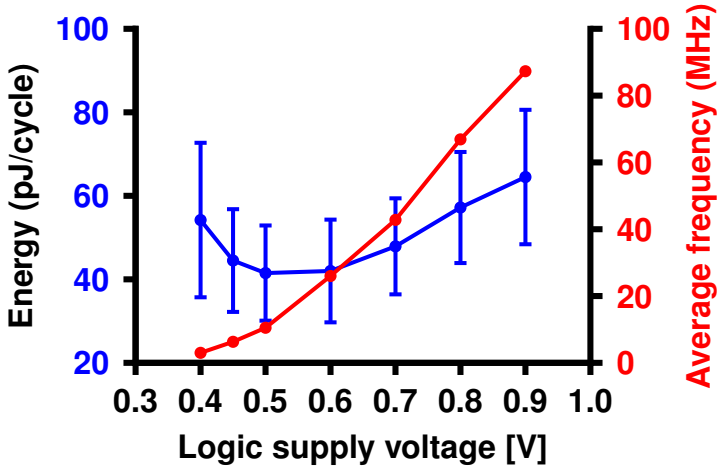


Figure 5.12. Multiple-die measurement from 2 different lots in the flat-mode.

direction. The minimum temperature achieved using the Peltier element is 5°C.

For the temperature measurement, the *MatMul* kernel is executed in the voltage-stacked mode with a supply voltage of 1.7 V. The kernel executes in CGRA-active mode at a frequency of 2.5 MHz. The voltage-stack is balanced with V_{MID} at 0.8 V and V_{BOT} at 0.4 V. Fig. 5.13 shows the variation of I_{in} ($\approx I_{TOP}$) and I_{MID} ($=I_{BOT}$) for temperature changing from 0°C to 80°C. Observed that, the SRAM current consumption increases at a faster rate as compared to that of the logic circuits with temperature variation. The system efficiency gradually decreases with an increase in temperature, as shown in Fig. 5.13. In our system, the system efficiency decreases from 95% to 92% for a temperature change from 0°C to 80°C. The impact of temperature change is significantly less as compared to the state-of-the-art system in [89]. In [89], the total system efficiency decreases from 84% down to 75% with an increase in temperature from -20°C to 70°C.

5.5 Comparison to the state-of-the-arts and discussions

Table 5.5 shows a comparison of our chip to state-of-the-art systems. Our design is the first voltage-stacked SoC with stacks operating in the near/sub-V_{th} region. The selected designs for this comparison are recent systems with logic circuit operating in

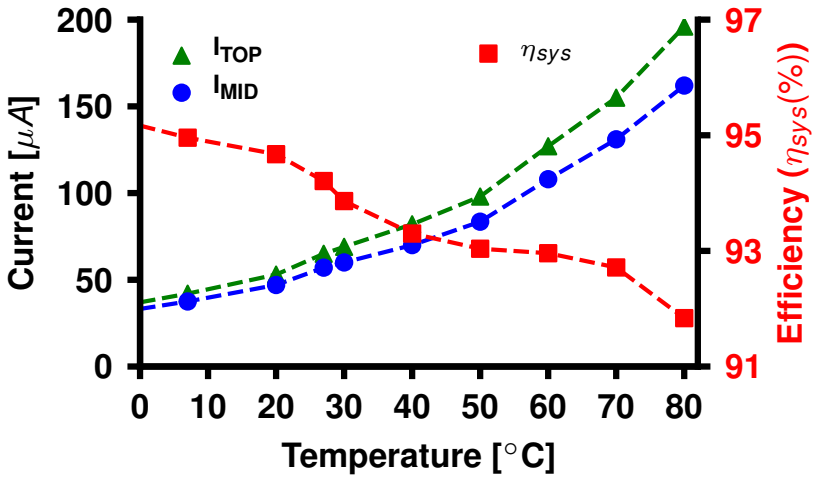


Figure 5.13. Measured I_{TOP}, I_{MID}, and system efficiency with temperature variations for the voltage-stacked system operating in the CGRA-active mode at 2.5 MHz.

the near/sub-V_{th} region. The designs in [61] and [64] are multi-core/accelerator based systems, designed using 28-nm FDSOI technology. The designs in [51] and [89] are processor systems with on-chip voltage regulators. The design in [49] is a processor system with an AES-128 accelerator and integrated power delivery. The design of [92] uses multiple SCVRs to convert the 2 V supply voltage to near/sub-V_{th}. Table 5.5 also shows efficiency, area overhead, energy efficiency, and energy consumption per cycle.

The key factors impacting the energy per cycle are system architecture, size and type of SRAM, and power delivery efficiency. As shown in Table 5.5, our system consists of foundry-provided 80 kB (6T) SRAM circuits. Note that other works are using low voltage custom memories (8T) to allow voltage scaling down to the near/sub-V_{th} region. These low voltage memories have relatively high area overhead when compared to 6T SRAMs and require extra custom design effort. The comparison shows that our system achieves the best total system efficiency (95%) with the least area overhead as compared to the other state-of-the-art works. In the state-of-the-art, it is unlikely to achieve >80% efficiency for a conversion ratio of <1/3. In [64], an 85% conversion efficiency is reported for output of 0.48 V generated from a 1.0 V supply with a conversion ratio of 1/2. In [89], a 0.4V is generated from a 1.55 V supply achieving ~80% efficiency. The conversion ratio in this work is 0.23 (1.7 V to 0.4 V),

we achieve >90% in most of the case as shown in Table 5.2.

The energy efficiency in [61] is better than our design. However, in [61], multiple external voltage supplies are required for SRAMs, logic circuits, IO-cells, and forward body biasing of the transistors. Our energy consumption per cycle is higher compared to the microprocessor systems in [51], [89], and [49]. However, the CGRA in our platform can perform up to 20 operations per cycle, while the processors in [51] and [89] have typically less than one operation/cycle. Although the measured energy consumption of our platform for the *MatMul* application in the CGRA-active mode is 39.2 pJ/cycle. The CGRA requires ∼24k cycles to execute *MatMul*. Therefore, the total energy required by our platform is 940 nJ. Alternatively, the RISC-V requires ∼134k cycles for *MatMul* application, assuming the energy consumption of 8 pJ/cycle [89]. The total energy required for the application would be 1072 nJ (14% higher compared to our platform). Hence, our system is better in terms of total energy consumption when compared with [51] and [89]. The area overhead of our controllers is $< 0.1\%$ of the total design area without considering the overhead of the voltage reference generators. Considering the voltage reference circuit in [147] designed in 28-nm FDSOI, the power consumption and area of the designed reference voltage generator (V_{ref}=∼0.9 V) are 100 nA (@1.2 V, 80°C) and 0.017 mm^2, respectively. For our design, the power consumption overhead would have been 0.01% while the estimated design area overhead would have been ∼1%. Additionally, in [149] and [150], the power consumption of the designed reference voltage generators is <300 pW. This would have resulted in negligible power consumption overhead.

As shown in Figs. 5.3a and 5.4, the measured voltage regulation by applying a step input on V_{MID} and V_{BOT} nodes is within 40 mV. This is an extreme use case. However, when running the benchmark applications in various operating modes, the voltage regulation of V_{MID} and V_{BOT} stays within 20 mV (2.5%) and 10 mV (2.5%), respectively, see Fig. 5.6. The designed SCVR in [64] can regulate within 90 mV (18%) for output of 0.48 V. In [47], the output voltage regulation is within 15 mV (4%) for an output of 0.4 V. In [142], the voltage regulation is within 30 mV (3%) for an output of 0.9 V on the voltage-stacked system operating at above-V_{th}, the intermediate rail is controlled using multiple small SCVRs.

The proposed voltage-stacked system shows resilience to process and temperature variations. The impact of process variations decreases from ∼40% to 33% for the voltage-stacked mode as compared to conventional flat-mode implementation. Furthermore, with an increase in temperature the system efficiency decreases by ∼4% as compared to the conventional system which decreases by ∼11% [89].

The current state-of-the-art voltage-stacking systems are designed for above-V_{th} operation, which cannot be scaled to operate in near/sub-V_{th} [111], [139]–[142]. Hence, these systems were not included in the comparison. Nevertheless, these

voltage-stacking systems achieve a total system efficiency of up to 96% with an area overhead of up to 30% for stack balancing using on-chip voltage converters.

5.6 Summary

The silicon measurement of the three-level voltage stacked system is presented in this chapter. The voltage-stack stability is proven by multiple stress benchmarks to force a current imbalance among the stacks. The charge recycling property of the voltage-stacks leads to a total system efficiency of up to 95%. The system efficiency is limited by the power consumption of the ABB controllers. The measured energy efficiency is 35.2 MMAC/s/mW with the chip operating at 1.7 V at 2.8 MHz. The energy efficiency is 1.6× higher on *average*, when compared to the conventional flat mode implementation. Additionally, the supply voltage for the voltage-stacked system is scaled down to 1.5 V to achieve minimum power consumption. The energy consumption at the minimum power point is 52% lower for the voltage-stacked system as compared to the flat-mode implementation. Furthermore, the measurement results show an improved resilience for process and temperature variations for the voltage-stacked system.

Table 5.5. Detailed comparison with state-of-the-art works.

	This work (Stacked)	**This work (Flat)**[††]	[64]	[51]	[138]	[61][†]	[89]	[49]	[92]
Circuit	RISC-V + CGRA		RISC-V + Vector Acc.	ARM CM0+	ARM CM0 + FFT Acc.	4× RISC-V core	RISC-V core	ARM CM0+ + AES-128	On-chip DC-DC VR
Region	Near/sub-V_{th}		Near/sub-V_{th}	Near/sub-V_{th}	Near/sub-V_{th}	Near/sub-V_{th}	Near/sub-V_{th}	Near/sub-V_{th}	Near/sub-V_{th}
Technology	28-nm FDSOI		28-nm FDSOI	65-nm	65-nm	28-nm FDSOI	28-nm FDSOI	65-nm	65-nm
Area	1.21 mm^2		1.07 mm^2	∼1 mm^2	0.47 mm^2	1.3 mm^2	0.06 mm^2	1.28 mm^2	
Area (overhead)	**300 μm^2 (<0.1%)**[*]	-	0.39 mm^2 (36%)	∼0.15 mm^2 (15%)	0.1 mm^2 (20%)	-	0.023 mm^2 (38%)	0.088 mm^2 (7%)	0.23 mm^2
SRAMs	80 kB (6T, foundry SRAM)		46 kB (8T)	4 kB (6T), 8 kB (8T)	512 B	32 kB (6T), 12 kB (SCM)	64 kB (6T), 8 kB (8T)	16 kB (6T), 8kB (10T)	-
Frequency	2.7 MHz	3.0 MHz	20 MHz	2.6 MHz	8.2 MHz	40.5 MHz	23 MHz	8 MHz	-
Supply voltage	1.7 V-0.8 V-0.4V	0.9V/0.4V	1.8 V, 1.0 V to 0.48V	1.2V to 0.35V	1.2V to 0.5V	0.46 V	1.5V to 0.3-0.5V	1.2 V to 0.48 V	2V to 0.4-0.7V
Converters	Controllers	External	SCVR	SCVR/LDO	SCVR	-	SCVR	SCVR/LDO	SCVR
η_{sys}	**95%**	-	82-89%	82%	82.4%	-	85%	82%	80%
Peak energy efficiency	35.2 MMACs/mW	20.5[††] MMACs/mW	20.9[#] MMACs/mW	-	-	95[†] MMACs/mW	-	-	-
Energy per cycle	39.2 pJ/cycle	67.5 pJ/cycle	60 pJ/cycle	12.4 pJ/cycle	50 pJ/cycle	20.7[†]pJ/cycle	8 pJ/cycle	23 pJ/cycle	-

[*] Controllers require 0.8V and 0.4V reference voltage generators (∼1% area overhead). Frequencies are the worst case achieved number.

[#] Double precision floating point vector accelerator, assuming 2 FLOP is 1 MAC.

[†] No on-chip converter for SRAMs, logic circuit, and FBB. The reported numbers are without conversion losses of voltage converters. We assume 2 OPs = 1 MAC.

[††] The flat-mode evaluation is performed assuming voltage conversion losses the same as depicted in Fig. 1.5.

Chapter 6

Conclusions and Future Perspectives

In this chapter, we make conclusions of this thesis. We first review the work that has been done in this thesis. Then, we present some prospective ideas and some open issues to extend this work.

6.1 Conclusions

Voltage scaling to the near/sub-threshold region is the most effective technique to reduce the energy consumption of a digital circuit. Typically, in the near/sub-threshold region, the performance degrades significantly along with the compromised robustness of the digital circuits. Moreover, the operation is highly susceptible to fabrication process variations and temperature. The past 20 years of near/sub-threshold design review in Chapter 1 highlighted the trends and improvements made in different directions. In this thesis, we further extend the state-of-the-art on the near/sub-threshold design concerning the challenges of process variations, performance degradation, and efficient power delivery to achieve an ultra-low power/energy digital design. This thesis introduces practical techniques to handle the complexity of low voltage design and enables robust and ultra-low energy systems with integrated power delivery having extremely low conversion loss and design/area overheads. This thesis focuses on realizing digital systems operating in near/sub-threshold regions with ultra-low energy consumption. In doing so, this thesis tackled the design challenges by advancing the field at four levels: the system, the architecture, the gate level, and the layout level.

In Chapter 2, an automatic flow for converting a flip-flop-based design to a latch-based design as well as a latch/flip-flop-mixed design is revealed. The proposed smart-retiming strategy shows insight into the timing and power consumption trade-offs for the latch-based and flip-flop-based designs. An optimum operating condition for the latch-based as well as the latch/flip-flop mixed design is identified using the

smart-retiming strategy for achieving the maximum time borrowing. The performance improvement for the latch-based design is up to 41% higher as compared to the flip-flop-based design in the super-threshold region of operation. In the near/sub-threshold region of operation, the performance improvement is up to 47%. The performance improvements are significant when the design is synthesized at an optimum operating frequency point. The gains of latch-based design diminish when the operating frequency is too high or too low. Furthermore, in the latch-based design, the power consumption distribution for the clock-tree increases from $\sim$ 16% to $\sim$ 40%, which reduces the gains of the latch-based design. In the super-threshold region, the performance enhancement is traded with supply voltage scaling, resulting in 21% and 16% power savings by the latch-based design and the mixed design, respectively, as compared to the flip-flop-based design in a 28-nm FDSOI CMOS technology.

In Chapter 3, a new standard cell library is developed for 8-track LVT 28-nm FDSOI CMOS technology for a near/sub-threshold operating supply voltage of 0.4 V. The designed standard cells are compatible with the existing foundry-provided library. The impact of process variation is reduced by balancing the pull-up network and pull-down network of each cell, resulting in the yield improvement by up to 14%. The experimental results by synthesizing ARM Cortex-M0 and various ITC benchmark circuits show a significant reduction of leakage power consumption by up to 50% without any frequency penalty and $\sim$1% area overhead compared to the foundry-provided PB0 library. Combining the newly developed library with the existing PB0 and PB4 libraries result in both leakage and dynamic power reduction, without any performance and area penalty. The synthesis result shows that the new design standard cell library is preferred over the foundry standard cell library. Furthermore, in Chapter 3, a qualitative standard cell library pruning methodology to detect and remove the cells from synthesis for near/sub-threshold operation is presented. The foundry provided library is often enough to meet the functional yield constraint up to a certain voltage in the sub-threshold region. The bad cells from the foundry library are detected based on the relative delay degradation of the cells when the voltage is scaled from nominal to near/sub-threshold. The proposed methodology even detects the cells which are allowed by the general guidelines provided in the literature for library pruning. The proposed library cell pruning methodology leads to delay spread reduction by 15% with process variations.

In chapter 4, a new methodology is proposed to tackle the power delivery-related issues for a design operating in the near/sub-threshold region. A three-level voltage-stacked system is designed with foundry-provided high-density SRAM blocks on the top stack and logic circuit operating in the near/sub-threshold in the middle and bottom domains, powered by a single 1.8 V (1.7 V$\pm$5%) external voltage source. The inherent voltage conversion property of voltage-stacking eliminates the requirement

of multiple voltage converters for SRAM and logic circuits. Additionally, the IO-cells of the designed chip is powered by the same external supply. The near/sub-threshold region operation of the stacks results in the balancing of the stacks with simpler circuits. The balancing of the intermediate rail between the stacks is done using current sink controllers and adaptive body-bias controllers. The leakage currents sourced by the SRAM blocks are sufficient to sustain the near/sub-threshold region operations of the logic circuits at the middle and bottom stacks. The adaptive body-bias voltage controller balances the MID and BOT stacks by changing the body-bias voltage to balance the current. Moreover, four different types of special level-shifters are designed for signals between stacks. The use of simpler circuitry results in system efficiency up to 95%.

In Chapter 5, the silicon measurement of the proposed three-level voltage stacked system is presented. The designed current sink and adaptive body-bias controllers are characterized in multiple stress configurations. The designed voltage controllers can balance the intermediate node voltage variation within a tolerance limit of $\pm 5\%$ of the voltage. The "converter-free" implementation results in significant area savings. The charge recycling property of the voltage-stacks leads to a total system efficiency of up to 95% which is the best in the state-of-the-art for any design operating in the near/sub-threshold region. The system efficiency is limited by the power consumption of the ABB controllers. The measured energy efficiency is 35.2 MMAC/s/mW with the chip operating at 1.7 V at 2.8 MHz. The energy efficiency is 1.6× higher on average when compared to the conventional flat mode implementation.

6.2 Future Perspectives

The near/sub-V_{th} design is highly promising for the emerging ultra-low-power applications in the consumer electronics market. Until now, voltage scaling to near/sub-V_{th} region is not common in the industry due to the various sub-V_{th} design challenges. Therefore, to meet the requirements for ultra-low-power applications, IC designers have explored various design techniques, design styles, power schemes as well as semiconductor IPs that have drastically increased the number of power domains and operating points in modern SoC. Over the past twenty years, significant research effort is dedicated to tackling the near/sub-V_{th} design problems and research is still ongoing to enable widespread adoption of near/sub-V_{th} circuits in the industry. Nowadays, to enable widespread adoption of sub-V_{th} design standard-cell libraries, IPs for compensating process, supply voltage, and temperature variations (PVT) to guarantee timing and power with a high yield are provided by some vendors [151], [152]. Additionally, a few IC vendors have started to use near/sub-V_{th} techniques in battery/energy harvesting operated applications [31], [32]. This thesis has advanced

the field of ultra-low-power digital system design to handle low voltage design complexity specifically related to performance, process variations, and efficient power delivery. Nevertheless, there are possible future directions that can further advance the state-of-the-art. The possible opportunities to extend this work are as follows:

- In Chapter 2, the presented latch-based design improves the performance significantly. However, the latch-based design shows that a significant amount of power is consumed by the clock-tree network to distribute the clock for the latches. The clock-tree network power consumption is $\sim 2\times$ as compared to that of the flip-flop-based design since the number of latches is $\geqslant 2\times$ the number of flip-flops (load capacitance for the clock-tree network is $\geqslant 2\times$). To address this significant increase in the clock-tree power, similar to multi-bit flip-flops, multi-bit latches can be used. As shown in [121], the use of multi-bit flip-flops can reduce the power consumption of the clock-tree network by up to 83%. Therefore, the use of multi-bit latches in the latch-based design can also improve the power consumption significantly while maintaining performance improvement because of time-borrowing properties. Furthermore, an additional direction of research for latch-based design is to incorporate the automatic clock-gating feature for the latches in the design.

- From the perspective of efficient power delivery, voltage stacking of power domains is a new trend that potentially could reduce the number of on-chip voltage converters. However, the possibility of enabling voltage-stacking alongside other low-power techniques opens several future directions of research. The fact that voltage-stacking recycles the current consumed in the top stack for powering the lower stacks, the technique like power gating requires further research in the architecture partitioning and intermediate rail balancing design techniques. From the perspective of design partitioning among stacks, a fine-grained approach alongside multiple parallel voltage-stacked designs integrated can be used to further improve the design flexibility.

- It is well known that the conversion efficiency of the on-chip voltage converters reduces significantly ($< 50\%$, see Fig. 1.6) when the load current is relatively small, for example, in the idle-mode, sleep-mode, or data-retention mode the current consumption is primarily the leakage current. The use of the voltage-stacking technique for such scenarios is an interesting future direction of research.

- In Chapter 5, the silicon measurement of the proposed voltage-stacked design reveals that the use of the current sink-based controller for the intermediate rails between the stacks limits the design flexibility in terms of increasing the

voltage for the logic stacks. This leads to an upper bound on the maximum frequency achieved by the presented voltage-stacked system. Therefore, an improved intermediate rail balancing circuitry with both current sink and source options will be an interesting improvement for the current voltage-stacked design.

- An additional advancement for the voltage-stacked design to improve the timing robustness is to have an on-chip adaptive clock generator that can adapt the operating frequency of the design based on the intermediate rail imbalance.

List of acronyms

ABB	adaptive body-biasing.
ADC	analog to digital converter.
AFE	Analog Front-End.
CDF	cumulative distribution function.
CGRA	coarse-grained reconfigurable array.
CPF	common power format.
CS	current sink.
DRC	design rule chek.
DSP	digital signal processor.
DUT	design under test.
EDA	electronic design automation.
EEG	electroencephalogram.
FBB	forward body-biasing.
FDSOI	Fully Depleted Silicon on Insulator.
FFT	fast Fourier transform.
FIR	finite impulse response.
FoG	freezing-of-gait.
FU	funtional unit.
IC	integrated circuit.
IoT	Internet-of-Things.

JLCC	J-Leaded Ceramic Chip Carrier.
LDO	low-dropout regulator.
LVT	low threshold voltage.
MAC	multiply-accumulate.
MatMul	matrix multiplication.
NLDM	non-linear delay model.
PB	poly-biasing.
PCB	printed circuit board.
PDN	pull-down network.
PUN	pull-up network.
RBB	reverse body-biasing.
RVT	regular threshold voltage.
SCVR	switched-capacitor voltage regulator.
SoC	system-on-chip.
SRAM	static random-access memory.
SVM	support-vector machine.
UTBB	Ultra-Thin Body and Buried Oxide.
VLIW	very long instruction word.

Bibliography

[1] *Ultra-low-power microcontroller market*, Jul. 2021. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/ultra-low-power-microcontroller-market-206772623.html.

[2] *Wearable brainwave processing platform*, Apr. 2021. [Online]. Available: http://brain-wave.nl/.

[3] *"Epilepsy." AANS*, Apr. 2021. [Online]. Available: https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Epilepsy.

[4] *"SUDEP", Epilepsy foundation*, May 2021. [Online]. Available: https://www.epilepsy.com/learn/early-death-and-sudep/sudep.

[5] *"Epilepsy and age." People with epilepsy | Expertise Centrum Kempenhaeghe*, Apr. 2021. [Online]. Available: https://www.kempenhaeghe.nl/epilepsie/people-with-epilepsy/.

[6] *EEG solutions." TMSi*, Apr. 2021. [Online]. Available: https://www.tmsi.com/products/saga32-64/.

[7] *"G.Nautilus Pro Wearable EEG Headset by G.tec." G.tec Medical engineering GmbH*, Apr. 2021. [Online]. Available: https://www.gtec.at/product/gnautilus-pro/.

[8] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," *Seizure*, vol. 40, pp. 88–101, 2016. DOI: 10.1016/j.seizure.2016.06.008.

[9] T. Roh, S. Hong, H. Cho, and H. Yoo, "A 259.6 μW nonlinear HRV-EEG chaos processor with body channel communication interface for mental health monitoring," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 294–296. DOI: 10.1109/ISSCC.2012.6177020.

[10] N. Verma, A. Shoeb, J. Bohorquez, J. Dawson, J. Guttag, and A. P. Chandrakasan, "A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 804–816, 2010. DOI: 10.1109/JSSC.2010.2042245.

[11] J. Hulzink, M. Konijnenburg, M. Ashouei, A. Breeschoten, T. Berset, J. Huisken, J. Stuyt, H. de Groot, F. Barat, J. David, and J. Van Ginderdeuren, "An ultra low energy biomedical signal processing system operating at near-threshold," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 6, pp. 546–554, 2011. DOI: 10.1109/TBCAS.2011.2176726.

[12] F. Rincón, J. Recas, N. Khaled, and D. Atienza, "Development and evaluation of multilead wavelet-based ECG delineation algorithms for embedded wireless sensor nodes," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 6, pp. 854–863, 2011. DOI: 10.1109/TITB.2011.2163943.

[13] W. Chen, H. Chiueh, T. Chen, C. Ho, C. Jeng, M. Ker, C. Lin, Y. Huang, C. Chou, T. Fan, M. Cheng, Y. Hsin, S. Liang, Y. Wang, F. Shaw, Y. Huang, C. Yang, and C. Wu, "A fully integrated 8-channel closed-loop neural-prosthetic CMOS SoC for real-time epileptic seizure control," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 232–247, 2014. DOI: 10.1109/JSSC.2013.2284346.

[14] J. Yoo, L. Yan, D. El-Damak, M. B. Altaf, A. Shoeb, H. Yoo, and A. Chandrakasan, "An 8-channel scalable EEG acquisition SoC with fully integrated patient-specific seizure classification and recording processor," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 292–294. DOI: 10.1109/ISSCC.2012.6177019.

[15] M. A. B. Altaf and J. Yoo, "A 1.83 μJ/classification, 8-channel, patient-specific epileptic seizure classification SoC using a non-linear support vector machine," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 1, pp. 49–60, 2016. DOI: 10.1109/TBCAS.2014.2386891.

[16] M. A. B. Altaf, C. Zhang, L. Radakovic, and J. Yoo, "Design of energy-efficient on-chip EEG classification and recording processors for wearable environments," in *IEEE International Symposium on Circuits and Systems*, May 2016. DOI: 10.1109/ISCAS.2016.7527443.

[17] Y. Wang, X. Long, H. V. Dijk, R. Aarts, and J. Arends, "Adaptive EEG channel selection for nonconvulsive seizure analysis," in *IEEE International Conference on Digital Signal Processing*, 2018, pp. 1–5. DOI: 10.1109/ICDSP.2018.8631844.

[18] B. de Bruin, K. Singh, J. Huisken, and H. Corporaal, "BrainWave: An energy-efficient EEG monitoring system - Evaluation and trade-offs," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020, pp. 181–186. DOI: 10.1145/3370748.3406571.

[19] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients," *Physica D: Nonlinear Phenomena*, vol. 144, no. 3, pp. 358–369, 2000. DOI: 10.1016/S0167-2789(00)00087-7.

[20] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated EEG analysis of epilepsy: A review," *Knowledge-Based Systems*, vol. 45, pp. 147–165, Jun. 2013. DOI: 10.1016/j.knosys.2013.02.014.

[21] L. Wang, J. Arends, X. Long, P. Cluitmans, and J. P. Dijk, "Seizure pattern-specific epileptic epoch detection in patients with intellectual disability," *Biomedical Signal Processing and Contro*, vol. 35, pp. 38–49, 2017. DOI: 10.1016/j.bspc.2017.02.008.

[22] J. Kwong and A. P. Chandrakasan, "An energy-efficient biomedical signal processing platform," *IEEE Journal of Solid-State Circuits*, vol. 46, Jul. 2011. DOI: 10.1109/JSSC.2011.2144450.

[23] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE Journal of Solid-State Circuits*, 2013. DOI: 10.1109/JSSC.2013.2253226.

[24] F. Montagna, S. Benatti, and D. Rossi, "Flexible, scalable and energy efficient bio-signals processing on the pulp platform: A case study on seizure detection," *Journal of Low Power Electronics and Applications*, vol. 7, no. 2, 2017. DOI: 10.3390/jlpea7020016.

[25] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, 2005. DOI: 10.1109/JSSC.2005.852162.

[26] D. Markovic, C. C. Wang, L. P. Alarcon, T. Liu, and J. M. Rabaey, "Ultralow-power design in near-threshold region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237–252, 2010. DOI: 10.1109/JPROC.2009.2035453.

[27] S. Jain, S. Khare, S. Yada, V. Ambili, P. Salihundam, S. Ramani, S. Muthukumar, M. Srinivasan, A. Kumar, S. Kumar, R. Ramanarayanan, V. Erraguntla, J. Howard, S. Vangal, S. Dighe, G. Ruhl, P. Aseron, H. Wilson, N. Borkar, V. De, and S. Borkar, "A 280mV-to-1.2V wide-operating-range IA-32 processor in 32nm CMOS," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 66–68. DOI: 10.1109/ISSCC.2012.6176932.

[28] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010. DOI: 10.1109/JPROC.2009.2034764.

[29] H. Kaul, M. Anders, S. Hsu, A. Agarwal, R. Krishnamurthy, and S. Borkar, "Near-threshold voltage (NTV) design: Opportunities and challenges," in *Proceedings of the ACM/IEEE Design Automation Conference*, 2012, pp. 1149–1154. DOI: 10.1145/2228360.2228572.

[30] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, 1997. DOI: 10.1109/4.604077.

[31] *Ambiq micro, inc.* Apr. 2021. [Online]. Available: http://www.ambiq.com/.

[32] *Everactive*, Apr. 2021. [Online]. Available: https://everactive.com/.

[33] A. Wang and A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," in *IEEE International Solid-State Circuits Conference*, 2004, 292–529 Vol.1. DOI: 10.1109/ISSCC.2004.1332709.

[34] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, D. Blaauw, and T. Austin, "A 2.60pJ/inst subthreshold sensor processor for optimal energy efficiency," in *Symposium on VLSI Circuits*, 2006, pp. 154–155. DOI: 10.1109/VLSIC.2006.1705356.

[35] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Performance and variability optimization strategies in a sub-200mV, 3.5pJ/inst, 11nW subthreshold processor," in *IEEE Symposium on VLSI Circuits*, 2007, pp. 152–153. DOI: 10.1109/VLSIC.2007.4342694.

[36] B. H. Calhoun and A. P. Chandrakasan, "A 256kb 65nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, 2007. DOI: 10.1145/2228360.2228572.

[37] N. Verma and A. P. Chandrakasan, "A 256kb 65nm 8T subthreshold SRAM employing sense-amplifier redundancy," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, 2008. DOI: 10.1109/JSSC.2007.908005.

[38] T. Kim, J. Liu, and C. H. Kim, "A voltage scalable 0.26 V, 64 kb 8T SRAM with V_{min} lowering techniques and deep sleep mode," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 6, pp. 1785–1795, 2009. DOI: 10.1109/CICC.2008.4672106.

[39] M. Chang, Y. Chiu, and W. Hwang, "Design and iso-area V_{min} analysis of 9T subthreshold SRAM with bit-interleaving scheme in 65-nm CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 7, pp. 429–433, 2012. DOI: 10.1109/TCSII.2012.2198984.

[40] S. Lutkemeier, T. Jungeblut, H. K. O. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65-nm 32b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 8–19, 2013. DOI: 10.1109/JSSC.2012.2220671.

[41] C. Lo and S. Huang, "P-P-N based 10T SRAM cell for low-leakage and resilient subthreshold operation," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 3, pp. 695–704, 2011. DOI: 10.1109/JSSC.2010.2102571.

[42] Y. Chiu, Y. Hu, M. Tu, J. Zhao, S. Jou, and C. Chuang, "A 40 nm 0.32 V 3.5 MHz 11T single-ended bit-interleaving subthreshold SRAM with data-aware write-assist," in *International Symposium on Low Power Electronics and Design*, 2013, pp. 51–56. DOI: 10.1109/ISLPED.2013.6629266.

[43] Y. Chiu, Y. Hu, M. Tu, J. Zhao, Y. Chu, S. Jou, and C. Chuang, "40 nm bit-interleaving 12T subthreshold SRAM with data-aware write-assist," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 9, pp. 2578–2585, 2014. DOI: 10.1109/TCSI.2014.2332267.

[44] J. Kwong, Y. K. Ramadass, N. Verma, and A. P. Chandrakasan, "A 65nm sub-V_t microcontroller with integrated SRAM and switched capacitor DC-DC converter," *IEEE Journal of Solid-State Circuits*, pp. 115–126, 2009. DOI: 10.1109/JSSC.2008.2007160.

[45] M. Seok, S. Hanson, Y. S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The Phoenix Processor: A 30pW platform for sensor applications," in *IEEE Symposium on VLSI Circuits*, 2008, pp. 188–189. DOI: 10.1109/VLSIC.2008.4586001.

[46] S. Lutkemeier, T. Jungeblut, M. Porrmann, and U. Rueckert, "A 200mV 32b subthreshold processor with adaptive supply voltage control," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 484–486. DOI: 10.1109/ISSCC.2012.6177101.

[47] D. Bol, J. De Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd, D. Flandre, and J. Legat, "SleepWalker: A 25MHz, 0.4V sub-mm^2 7μW/MHz microcontroller in 65nm LP/GP CMOS for low-carbon wireless sensor nodes," *IEEE Journal of Solid-State Circuits*, 2013. DOI: 10.1109/JSSC.2012.2218067.

[48] M. Fojtik, D. Kim, G. Chen, Y. Lin, D. Fick, J. Park, M. Seok, M. Chen, Z. Foo, D. Blaauw, and D. Sylvester, "A millimeter-scale energy-autonomous sensor system with stacked battery and solar cells," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 801–813, 2013. DOI: 10.1109/JSSC.2012.2233352.

[49] J. Myers, A. Savanth, D. Howard, R. Gaddh, P. Prabhat, and D. Flynn, "An 80nW retention 11.7pJ/cycle active subthreshold ARM Cortex-M0+ subsystem in 65nm CMOS for WSN applications," in *IEEE International Solid-State Circuits Conference*, 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7062967.

[50] S. Clerc, M. Saligane, F. Abouzeid, M. Cochet, J.-M. Daveau, C. Bottoni, D. Bol, J. De-Vos, D. Zamora, B. Coeffic, D. Soussan, D. Croain, M. Naceur, P. Schamberger, P. Roche, and D. Sylvester, "A 0.33V/-40°C process/temperature closed-loop compensation SoC embedding all-digital clock multiplier and DC-DC converter exploiting FDSOI 28nm back-gate biasing," in *IEEE International Solid-State Circuits Conference*, 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7062970.

[51] J. Myers, A. Savanth, P. Prabhat, S. Yang, R. Gaddh, S. O. Toh, and D. Flynn, "A 12.4pJ/cycle sub-threshold, 16pJ/cycle near-threshold ARM Cortex-M0+ MCU with autonomous SRPG/DVFS and temperature tracking clocks," in *Symposium on VLSI Circuits*, 2017. DOI: 10.23919/VLSIC.2017.8008529.

[52] H. Reyserhove and W. Dehaene, "A differential transmission gate design flow for minimum energy sub-10pJ/cycle ARM Cortex-M0 MCUs," *IEEE Journal of Solid-State Circuits*, pp. 1904–1914, 2017. DOI: 10.1109/JSSC.2017.2693241.

[53] S. Paul, V. Honkote, R. G. Kim, T. Majumder, P. A. Aseron, V. Grossnickle, R. Sankman, D. Mallik, T. Wang, S. Vangal, J. W. Tschanz, and V. De, "A sub-cm^3 energy-harvesting stacked wireless sensor node featuring a near-threshold voltage IA-32 microcontroller in 14-nm tri-gate CMOS for always-on always-sensing applications," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 961–971, 2017. DOI: 10.1109/JSSC.2016.2638465.

[54] R. Uytterhoeven and W. Dehaene, "A sub 10 pJ/cycle over a 2 to 200 MHz performance range RISC- V microprocessor in 28nm FDSOI," *European Solid*

State Circuits Conference, pp. 236–239, 2018. DOI: 10.1109/ESSCIRC.2018.8494259.

[55] G. Lallement, F. Abouzeid, M. Cochet, J. Daveau, P. Roche, and J. Autran, "A 2.7 pJ/cycle 16 MHz, 0.7 μW deep sleep power ARM Cortex-M0+ core SoC in 28 nm FDSOI," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 7, pp. 2088–2100, 2018. DOI: 10.1109/JSSC.2018.2821167.

[56] R. Salvador, A. Sanchez, X. Fan, and T. Gemmeke, "A Cortex-M3 based MCV featuring AVS with 34nW static power, 15.3pJ/inst. active energy, and 16% power variation across process and temperature," *IEEE European Solid State Circuits Conference*, pp. 278–281, 2018. DOI: 10.1109/ESSCIRC.2018.8494312.

[57] D. Bol, M. Schramme, L. Moreau, T. Haine, P. Xu, C. Frenkel, R. Dekimpe, F. Stas, and D. Flandre, "A 40-to-80MHz sub-4μW/MHz ULV Cortex-M0 MCU SoC in 28nm FDSOI with dual-loop adaptive back-bias generator for 20μs wake-up from deep fully retentive sleep mode," in *IEEE International Solid- State Circuits Conference*, 2019, pp. 322–324. DOI: 10.1109/ISSCC.2019.8662293.

[58] M. Pons, C. Müller, D. Ruffieux, J.-L. Nagel, S. Emery, A. Burg, S. Tanahashi, Y. Tanaka, and A. Takeuchi, "A 0.5V 2.5μW/MHz microcontroller with analog-assisted adaptive body bias PVT compensation with 3.13nW/kB SRAM retention in 55nm deeply-depleted channel CMOS," in *IEEE Custom Integrated Circuits Conference*, 2019, pp. 1–4. DOI: 10.1109/CICC.2019.8780199.

[59] J. Lee, Y. Zhang, Q. Dong, W. Lim, M. Saligane, Y. Kim, S. Jeong, J. Lim, M. Yasuda, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, "A 6.4pJ/cycle self-tuning Cortex-M0 IoT processor based on leakage-ratio measurement for energy-optimal operation across wide-range PVT variation," in *IEEE International Solid- State Circuits Conference*, 2019, pp. 314–315. DOI: 10.1109/ISSCC.2019.8662454.

[60] P. Prabhat, B. Labbe, G. Knight, A. Savanth, J. Svedas, M. J. Walker, S. Jeloka, P. M.-Y. Fan, F. Garcia-Redondo, T. Achuthan, and J. Myers, "M0N0: A performance-regulated 0.8-to-38MHz DVFS ARM Cortex-M33 SIMD MCU with 10nW sleep power," in *IEEE International Solid- State Circuits Conference*, 2020, pp. 422–424. DOI: 10.1109/ISSCC19947.2020.9063136.

[61] D. Rossi, A. Pullini, I. Loi, M. Gautschi, F. K. Gürkaynak, A. Teman, J. Constantin, A. Burg, I. Panades, E. Beigné, F. Clermidy, F. Abouzeid, P. Flatresse, and L. Benini, "193 MOPS/mW, 162MOPS, 0.32V to 1.15V voltage

range multi-core accelerator for energy efficient parallel and sequential digital processing," in *IEEE Symposium in Low-Power and High-Speed Chips*, 2016, pp. 1–3. DOI: 10.1109/CoolChips.2016.7503670.

[62] C. Kim, M. Chung, Y. Cho, M. Konijnenburg, S. Ryu, and J. Kim, "ULP-SRP: Ultra low-power Samsung reconfigurable processor for biomedical applications," *ACM TRETS*, 2014. DOI: 10.1145/2629610.

[63] D. Rossi, A. Pullini, I. Loi, M. Gautschi, F. K. Gürkaynak, A. Bartolini, P. Flatresse, and L. Benini, "A 60 GOPS/W, −1.8V to 0.9V body bias ULP cluster in 28nm UTBB FD-SOI technology," *Solid-State Electronics*, vol. 117, pp. 170–184, 2016. DOI: 10.1016/j.sse.2015.11.015.

[64] B. Keller, M. Cochet, B. Zimmer, J. Kwak, A. Puggelli, Y. Lee, M. Blagojević, S. Bailey, P.-F. Chiu, P. Dabbelt, C. Schmidt, E. Alon, K. Asanović, and B. Nikolić, "A RISC-V processor SoC with integrated power management at submicrosecond timescales in 28nm FDSOI," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1863–1875, 2017. DOI: 10.1109/JSSC.2017.2690859.

[65] F. Abouzeid, S. Clerc, C. Bottoni, B. Coeffic, J.-M. Daveau, D. Croain, G. Gasiot, D. Soussan, and P. Roche, "28nm FD-SOI technology and design platform for sub-10pJ/cycle and SER-immune 32bits processors," in *IEEE European Solid-State Circuits Conference*, 2015, pp. 108–111. DOI: 10.1109/ESSCIRC.2015.7313840.

[66] M. Ashouei, J. Hulzink, M. Konijnenburg, J. Zhou, F. Duarte, A. Breeschoten, J. Huisken, J. Stuyt, H. de Groot, F. Barat, J. David, and J. Van Ginderdeuren, "A voltage-scalable biomedical signal processor running ecg using 13pj/cycle at 1mhz and 0.4v," in *IEEE International Solid-State Circuits Conference*, 2011, pp. 332–334. DOI: 10.1109/ISSCC.2011.5746341.

[67] A. Akram and L. Sawalha, "A survey of computer architecture simulation techniques and tools," *IEEE Access*, vol. 7, pp. 78 120–78 145, 2019. DOI: 10.1109/ACCESS.2019.2917698.

[68] Y. Pu, J. Pineda de Gyvez, H. Corporaal, and Y. Ha, "An ultra-low-energy multi-standard JPEG co-processor in 65 nm CMOS with sub/near threshold supply voltage," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 3, pp. 668–680, 2010. DOI: 10.1109/JSSC.2009.2039684.

[69] A. Quelen, G. Pillonnet, P. Flatresse, and E. Beigné, "A 2.5μW 0.0067mm^2 automatic back-biasing compensation unit achieving 50% leakage reduction in FDSOI 28nm over 0.35-to-1V VDD range," in *IEEE International Solid - State Circuits Conference*, 2018, pp. 304–306. DOI: 10.1109/ISSCC.2018.8310305.

[70] G. Lallement, F. Abouzeid, J. Daveau, P. Roche, and J. Autran, "A 1.1pJ/cycle, 20MHz, 0.42V temperature compensated ARM Cortex-M0+ SoC with adaptive self body-biasing in FDSOI," *IEEE Solid-State Circuits Letters*, vol. 1, no. 7, pp. 174–177, 2018. DOI: 10.1109/LSSC.2019.2897016.

[71] M. Saligane, J. Lee, Q. Dong, M. Yasuda, K. Kumeno, F. Ohno, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, "An adaptive body-biasing SoC using in situ slack monitoring for runtime replica calibration," in *IEEE Symposium on VLSI Circuits*, 2018, pp. 63–64. DOI: 10.1109/VLSIC.2018.8502411.

[72] S. Höppner, H. Eisenreich, D. Walter, A. Scharfe, A. Oefelein, F. Schraut, J. Schreiter, T. Riedel, H. Bauer, R. Niebsch, S. Scherzer, T. Hocker, S. Scholze, S. Henker, M. Nossmann, U. Hensel, and H. Prengel, "Adaptive body bias aware implementation for ultra-low-voltage designs in 22FDX technology," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 10, pp. 2159–2163, 2020. DOI: 10.1109/TCSII.2019.2959544.

[73] M. Konijnenburg, Y. Cho, M. Ashouei, T. Gemmeke, C. Kim, J. Hulzink, J. Stuyt, M. Jung, J. Huisken, S. Ryu, J. Kim, and H. de Groot, "Reliable and energy-efficient 1MHz 0.4V dynamically reconfigurable SoC for ExG applications in 40nm LP CMOS," in *IEEE International Solid-State Circuits Conference*, 2013, pp. 430–431. DOI: 10.1109/ISSCC.2013.6487801.

[74] A. Gebregiorgis, M. S. Golanbari, S. Kiamehr, F. Oboril, and M. B. Tahoori, "Maximizing energy efficiency in NTC by variation-aware microprocessor pipeline optimization," in *ACM/IEEE International Symposium on Low Power Electronics and Design*, San Francisco Airport, CA, USA, 2016, pp. 272–277. DOI: 10.1145/2934583.2934635.

[75] D. Jeon, M. Seok, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A superpipelined energy efficient subthreshold 240 MS/s FFT core in 65nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 23–34, 2012. DOI: 10.1109/JSSC.2011.2169311.

[76] B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *ACM/IEEE International Symposium on Low Power Electronics and Design*, 2007. DOI: 10.1145/1283780.1283789.

[77] N. Reynders and W. Dehaene, "A 210mV 5MHz variation-resilient near-threshold JPEG encoder in 40nm CMOS," in *IEEE International Solid-State Circuits Conference*, 2014, pp. 456–457. DOI: 10.1109/ISSCC.2014.6757511.

[78] H. Mostafa, M. Anis, and M. Elmasry, "A novel low area overhead direct adaptive body bias (D-ABB) circuit for die-to-die and within-die variations compensation," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 19, no. 10, pp. 1848–1860, 2011. DOI: 10.1109/TVLSI.2010.2060503.

[79] N. Kamae, A. K. M. M. Islam, A. Tsuchiya, and H. Onodera, "A body bias generator with wide supply-range down to threshold voltage for within-die variability compensation," in *IEEE Asian Solid-State Circuits Conference*, 2014, pp. 53–56. DOI: 10.1109/ASSCC.2014.7008858.

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

[81] X. Wang, M. Tehranipoor, S. George, D. Tran, and L. Winemberg, "Design and analysis of a delay sensor applicable to process/environmental variations and aging measurements," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 20, no. 8, pp. 1405–1418, 2012. DOI: 10.1109/TVLSI.2011.2158124.

[82] A. K. M. M. Islam and H. Onodera, "On-chip monitoring and compensation scheme with fine-grain body biasing for robust and energy-efficient operations," in *Asia and South Pacific Design Automation Conference*, 2016, pp. 403–409. DOI: 10.1109/ASPDAC.2016.7428045.

[83] H. Ahmadi Balef, H. Fatemi, K. Goossens, and J. Pineda de Gyvez, "Effective in-situ chip health monitoring with selective monitor insertion along timing paths," in *Proceedings of the Great Lakes Symposium on VLSI*, Chicago, IL, USA, 2018, pp. 213–218. DOI: 10.1145/3194554.3194563.

[84] F. U. Rahman, R. Pamula, A. Boora, X. Sun, and V. Sathe, "Computationally enabled total energy minimization under performance requirements for a voltage-regulated 0.38-to-0.58V microprocessor in 65nm CMOS," in *IEEE International Solid- State Circuits Conference*, 2019. DOI: 10.1109/ISSCC.2019.8662486.

[85] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: A low-power pipeline based on circuit-level timing speculation," in *IEEE/ACM International Symposium on Microarchitecture,*, 2003. DOI: 10.1109/MICRO.2003.1253179.

[86] Y. Zhang, M. Khayatzadeh, K. Yang, M. Saligane, N. Pinckney, M. Alioto, D. Blaauw, and D. Sylvester, "Irazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor," in *IEEE International Solid-State Circuits Conference*, 2016. DOI: 10.1109/ISSCC.2016.7417956.

[87] H. Ahmadi Balef, H. Fatemi, K. Goossens, and J. Pineda De Gyvez, "Timing speculation with optimal in situ monitoring placement and within-cycle error prevention," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 27, no. 5, pp. 1206–1217, 2019. DOI: 10.1109/TVLSI.2019.2895972.

[88] Y. Zhao, Y. Yang, K. Mazumdar, X. Guo, and M. R. Stan, "A multi-output on-chip switched-capacitor DC-DC converter for near- and sub-threshold power modes," in *IEEE International Symposium on Circuits and Systems*, 2014, pp. 1632–1635. DOI: 10.1109/ISCAS.2014.6865464.

[89] M. Turnquist, M. Hiienkari, J. Mäkipää, R. Jevtic, E. Pohjalainen, T. Kallio, and L. Koskinen, "Fully integrated DC-DC converter and a 0.4V 32-bit CPU with timing-error prevention supplied from a prototype 1.55V Li-ion battery," in *Symposium on VLSI Circuits*, 2015. DOI: 10.1109/VLSIC.2015.7231307.

[90] M. Kutila, A. Paasio, and T. Lehtonen, "Comparison of 130nm technology 6T and 8T sram cell designs for near-threshold operation," in *IEEE International Midwest Symposium on Circuits and Systems*, 2014. DOI: 10.1109/MWSCAS.2014.6908567.

[91] L. G. Salem and P. P. Mercier, "An 85%-efficiency fully integrated 15-ratio recursive switched-capacitor DC-DC converter with 0.1-to-2.2V output voltage range," *IEEE International Solid-State Circuits Conference*, pp. 88–89, 2014. DOI: 10.1109/ISSCC.2014.6757350.

[92] J. Jiang, Y. Lu, C. Huang, W. Ki, and P. K. T. Mok, "A 2-/3-phase fully integrated switched-capacitor DC-DC converter in bulk CMOS for energy-efficient digital circuits with 14% efficiency improvement," in *IEEE International Solid-State Circuits Conference*, 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7063078.

[93] T. M. Andersen, F. Krismer, J. W. Kolar, T. Toifl, C. Menolfi, L. Kull, T. Morf, M. Kossel, M. Brändli, and P. A. Francese, "A 10W on-chip switched capacitor voltage regulator with feedforward regulation capability for granular microprocessor power delivery," *IEEE Transactions on Power Electronics*, vol. 32, no. 1, pp. 378–393, 2017. DOI: 10.1109/TPEL.2016.2530745.

[94] Y. K. Ramadass and A. P. Chandrakasan, "Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications," in *IEEE Power Electronics Specialists Conference*, 2007, pp. 2353–2359. DOI: 10.1109/PESC.2007.4342378.

[95] J. De Vos, D. Flandre, and D. Bol, "A sizing methodology for on-chip switched-capacitor DC/DC converters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 5, pp. 1597–1606, 2014. DOI: 10.1109/TCSI.2013.2285692.

[96] B. Zimmer, Y. Lee, A. Puggelli, J. Kwak, R. Jevtić, B. Keller, S. Bailey, M. Blagojević, P. Chiu, H. Le, P. Chen, N. Sutardja, R. Avizienis, A. Waterman, B. Richards, P. Flatresse, E. Alon, K. Asanović, and B. Nikolić, "A RISC-V vector processor with simultaneous-switching switched-capacitor DC–DC converters in 28nm FDSOI," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 930–942, 2016. DOI: 10.1109/JSSC.2016.2519386.

[97] J. De Vos, D. Flandre, and D. Bol, "A dual-mode DC-DC converter for ultra-low-voltage microcontrollers," in *IEEE Subthreshold Microelectronics Conference*, 2012, pp. 1–3. DOI: 10.1109/SubVT.2012.6404306.

[98] H. Le, S. R. Sanders, and E. Alon, "Design techniques for fully integrated switched-capacitor dc-dc converters," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 9, pp. 2120–2131, 2011. DOI: 10.1109/JSSC.2011.2159054.

[99] G. A. Rincón-Mora and P. Allen, "A low-voltage, low quiescent current, low drop-out regulator," *IEEE Journal of Solid-state Circuits*, vol. 33, pp. 36–44, 1998. DOI: 10.1109/4.654935.

[100] Y. K. Ramadass and A. P. Chandrakasan, "Minimum energy tracking loop with embedded DC–DC converter enabling ultra-low-voltage operation down to 250 mV in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 256–265, 2008. DOI: 10.1109/JSSC.2007.914720.

[101] Y. Okuma, K. Ishida, Y. Ryu, X. Zhang, P.-H. Chen, K. Watanabe, M. Takamiya, and T. Sakurai, "0.5V input digital LDO with 98.7% current efficiency and 2.7-μA quiescent current in 65nm CMOS," in *IEEE Custom Integrated Circuits Conference*, 2010, pp. 1–4. DOI: 10.1109/CICC.2010.5617586.

[102] K. Yoshikawa, K. Kanamaru, S. Inui, Y. Hagihara, Y. Nakamura, and T. Yoshimura, "Timing optimization by replacing flip-flops to latches," in *Proceedings of the IEEE Asia and South Pacific Design Automation Conference*, Jan. 2004, pp. 186–191. DOI: 10.1109/ASPDAC.2004.1337563.

[103] Y. Hagihara, S. Inui, A. Yoshikawa, T. Uesugi, T. Osada, S. Nakazato, M. Ikeda, M. Okada, and S. Yamada, "A skew-tolerant design scheme for over 1-GHz LSIs," in *Proceedings of the IEEE European Solid-State Circuits Conference*, Sep. 2000, pp. 415–418. DOI: 10.1109/ESSCIR.2000.186539.

[104] T. Baumann, D. Schmitt-Landsiedel, and C. Pacha, "Architectural assessment of design techniques to improve speed and robustness in embedded microprocessors," in *Proceedings of the ACM/IEEE Design Automation Conference*, Jul. 2009, pp. 947–950. DOI: 10.1145/1629911.1630154.

[105] D. Chinnery, K. Keutzer, J. Sanghavi, E. Killian, and K. Sheth, "Closing the gap between asic and custom: Tools and techniques for high-performance asic design," in. Kluwer New York, 2004.

[106] T.-Y. Wu and Y.-L. Lin, "Storage optimization by replacing some flip-flops with latches," in *Proceedings of the IEEE European Design Automation Conference*, Sep. 1996, pp. 296–301. DOI: 10.1109/EURDAC.1996.558220.

[107] Y. Zhang and B. H. Calhoun, "Hold time closure for subthreshold circuits using a two-phase, latch based timing method," in *Proceedings of the IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, Oct. 2013, pp. 1–2. DOI: 10.1109/S3S.2013.6716531.

[108] A. Roy and B. H. Calhoun, "Exploring circuit robustness to power supply variation in low-voltage latch and register-based digital systems," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 2016, pp. 273–276. DOI: 10.1109/ISCAS.2016.7527223.

[109] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. M. Harris, D. Blaauw, and D. Sylvester, "Bubble razor: Eliminating timing margins in an arm cortex-m3 processor in 45 nm cmos using architecturally independent error detection and correction," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 66–81, Jan. 2013. DOI: 10.1109/JSSC.2012.2220912.

[110] M. Pons, T. C. Le, C. Arm, D. Séverac, J. L. Nagel, M. Morgan, and S. Emery, "Sub-threshold latch-based icyflex2 32-bit processor with wide supply range operation," in *Proceedings of the IEEE European Solid-State Circuits Conference*, Sep. 2016, pp. 41–44. DOI: 10.1109/ESSCIRC.2016.7598238.

[111] S. Rajapandian, Zheng Xu, and K. L. Shepard, "Implicit DC-DC down conversion through charge-recycling," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 846–852, 2005. DOI: 10.1109/JSSC.2004.842861.

[112] R. Trapani Possignolo, E. Ebrahimi, E. K. Ardestani, A. Sankaranarayanan, J. L. Briz, and J. Renau, "GPU NTC process variation compensation with voltage stacking," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 26, no. 9, pp. 1713–1726, 2018. DOI: 10.1109/TVLSI.2018.2831665.

[113] D. Chinnery and K. Keutzer, "Improving performance through microarchitecture," in. Jan. 2004, pp. 33–56. DOI: 10.1007/0-306-47823-4_2.

[114] S. Kim, I. Han, S. Paik, and Y. Shin, "Pulser gating: A clock gating of pulsed-latch circuits," in *Proceedings of the IEEE Asia and South Pacific Design Automation Conference*, Jan. 2011, pp. 190–195. DOI: 10.1109/ASPDAC.2011.5722182.

[115] S. Paik, G. J. Nam, and Y. Shin, "Implementation of pulsed-latch and pulsed-register circuits to minimize clocking power," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2011, pp. 640–646. DOI: 10.1109/ICCAD.2011.6105397.

[116] A. Chandrakasan, W. J. Bowhill, and F. Fox, "Circuit styles for logics," in *Design of High-Performance Microprocessor Circuits*. Wiley-IEEE Press, 2001.

[117] D. Chinnery and K. Keutzer, "Reducing the timing overhead," in. Jan. 2004, pp. 57–100. DOI: 10.1007/0-306-47823-4_3.

[118] J. Leijten, J. van Meerbergen, and J. Jess, "Analysis and reduction of glitches in synchronous networks," in *Proceedings of the IEEE European Design and Test Conference.*, Mar. 1995, pp. 398–403. DOI: 10.1109/EDTC.1995.470365.

[119] M. Wieckowski, Young Min Park, C. Tokunaga, Dong Woon Kim, Zhiyoong Foo, D. Sylvester, and D. Blaauw, "Timing yield enhancement through soft edge flip-flop based design," in *IEEE Custom Integrated Circuits Conference*, 2008, pp. 543–546. DOI: 10.1109/CICC.2008.4672142.

[120] Y. Hwang, J. Lin, and M. Sheu, "Low-power pulse-triggered flip-flop design with conditional pulse-enhancement scheme," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 20, no. 2, pp. 361–366, 2012. DOI: 10.1109/TVLSI.2010.2096483.

[121] K. Singh, O. A. R. Rosas, H. Jiao, J. Huisken, and J. P. de Gyvez, "Multi-bit pulsed-latch based low power synchronous circuit design," in *IEEE International Symposium on Circuits and Systems*, 2018, pp. 1–5. DOI: 10.1109/ISCAS.2018.8351251.

[122] P. Magarshack, P. Flatresse, and G. Cesana, "UTBB FD-SOI: A process/design symbiosis for breakthrough energy-efficiency," in *Design, Automation Test in Europe Conference Exhibition*, Mar. 2013, pp. 952–957. DOI: 10.7873/DATE.2013.200.

[123] P. Flatresse, B. Giraud, J. Noel, B. Pelloux-Prayer, F. Giner, D. Arora, F. Arnaud, N. Planes, J. Le Coz, O. Thomas, S. Engels, G. Cesana, R. Wilson, and P. Urard, "Ultra-wide body-bias range LDPC decoder in 28-nm UTBB FDSOI technology," in *IEEE International Solid-State Circuits Conference*, 2013, pp. 424–425. DOI: 10.1109/ISSCC.2013.6487798.

[124] S. A. Vitale, P. W. Wyatt, N. Checka, J. Kedzierski, and C. L. Keast, "FD-SOI process technology for subthreshold-operation ultralow-power electronics," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 333–342, Feb. 2010. DOI: 10.1109/JPROC.2009.2034476.

[125] A. A. Vatanjou, E. Late, T. Ytterdal, and S. Aunet, "Ultra-low voltage adders in 28 nm FDSOI exploring poly-biasing for device sizing," in *IEEE Nordic Circuits and Systems Conference*, Nov. 2016, pp. 1–4. DOI: 10.1109/NORCHIP.2016.7792895.

[126] H. Cai, Y. Wang, L. A. De Barros Naviner, and W. Zhao, "Robust ultra-low power non-volatile logic-in-memory circuits in FD-SOI technology," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 4, pp. 847–857, Apr. 2017. DOI: 10.1109/TCSI.2016.2621344.

[127] F. Abouzeid, S. Clerc, B. Pelloux-Prayer, F. Argoud, and P. Roche, "28-nm CMOS, energy efficient and variability tolerant, 350mV-to-1.0V, 10MHz/700MHz, 252bits frame error-decoder," in *Proceedings of the ESSCIRC*, 2012, pp. 153–156. DOI: 10.1109/ESSCIRC.2012.6341282.

[128] S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, and A. Fish, "Digital subthreshold logic design - motivation and challenges," in *IEEE Convention of Electrical and Electronics Engineers*, 2008. DOI: 10.1109/EEEI.2008.4736624.

[129] W. T. Wong, K. Singh, J. Huisken, and J. P. de Gyvez, "Power and variation improved near-V_{th} standard cell library for 28-nm FDSOI," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2019. DOI: 10.1109/S3S46989.2019.9320687.

[130] J. Keane, H. Eom, T. Kim, S. Sapatnekar, and C. Kim, "Subthreshold logical effort: A systematic framework for optimal subthreshold device sizing," in *ACM/IEEE Design Automation Conference*, 2006. DOI: 10.1145/1146909.1147022.

[131] H. Soeleman, K. Roy, and B. C. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 9, no. 1, pp. 90–99, 2001. DOI: 10.1109/92.920822.

[132] Y. Pu, J. P. de Gyvez, H. Corporaal, and Y. Ha, "Vt balancing and device sizing towards high yield of sub-threshold static logic gates," in *ACM/IEEE International Symposium on Low Power Electronics and Design*, 2007, pp. 355–358. DOI: 10.1145/1283780.1283857.

[133] B. Liu, M. Ashouei, J. Huisken, and J. P. de Gyvez, "Standard cell sizing for subthreshold operation," in *Design Automation Conference*, 2012, pp. 962–967. DOI: 10.1145/2228360.2228533.

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

[135] M. Alioto, E. Consoli, and G. Palumbo, "General strategies to design nanometer flip-flops in the energy-delay space," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1583–1596, Jul. 2010. DOI: 10.1109/TCSI.2009.2033538.

[136] D. Markovic, B. Nikolic, and R. W. Brodersen, "Analysis and design of low-energy flip-flops," in *Proceedings of the International Symposium on Low Power Electronics and Design*, Aug. 2001, pp. 52–55. DOI: 10.1109/LPE.2001.945371.

[137] A. Dendouga, N. Bouguechal, S. Barra, and O. Manck, "Timing characterization and layout of a low power differential C2MOS flip flop in 0.35μm technology," in *2008 2nd International Conference on Signals, Circuits and Systems*, Nov. 2008, pp. 1–4. DOI: 10.1109/ICSCS.2008.4746891.

[138] F. U. Rahman, R. Pamula, and V. S. Sathe, "Computationally enabled minimum total energy tracking for a performance regulated sub-threshold microprocessor in 65nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 2, pp. 494–504, 2020. DOI: 10.1109/JSSC.2019.2956884.

[139] S. K. Lee, T. Tong, X. Zhang, D. Brooks, and G. Y. Wei, "A 16-Core voltage-stacked system with adaptive clocking and an integrated switched-capacitor DC-DC converter," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 25, no. 4, pp. 1271–1284, 2017. DOI: 10.1109/TVLSI.2016.2633805.

[140] A. Zou, J. Leng, X. He, Y. Zu, C. D. Gill, V. Janapa Reddi, and X. Zhang, "Voltage-stacked GPUs: A control theory driven cross-layer solution for practical voltage stacking in GPUs," *Proceedings of the Annual International Symposium on Micro-architecture*, vol. 2018-October, pp. 390–402, 2018. DOI: 10.1109/MICRO.2018.00039.

[141] K. Ueda, F. Morishita, S. Okura, L. Okamura, T. Yoshihara, and K. Arimoto, "Low-power on-chip charge-recycling DC-DC conversion circuit and system," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2608–2617, 2013. DOI: 10.1109/JSSC.2013.2274829.

[142] K. Blutman, A. Kapoor, A. Majumdar, J. G. Martinez, J. Echeverri, L. Sevat, A. P. Van Der Wel, H. Fatemi, K. A. Makinwa, and J. P. De Gyvez, "A low-power microcontroller in a 40-nm CMOS using charge recycling," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 950–960, 2017. DOI: 10.1109/JSSC.2016.2637003.

[143] L. Duch, S. Basu, R. Braojos, G. Ansaloni, L. Pozzi, and D. Atienza, "HEALWEAR: An ultra-low power heterogeneous system for bio-signal analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2017. DOI: 10.1109/TCSI.2017.2701499.

[144] S. Das, K. J. M. Martin, D. Rossi, P. Coussy, and L. Benini, "An energy-efficient integrated programmable array accelerator and compilation flow for near-sensor ultra low power processing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2019. DOI: 10.1109/TCAD.2018.2834397.

[145] A. Traber, F. Zaruba, S. Stucki, A. Pullini, G. Haugou, E. Flamand, F. K. Gurkaynak, and L. Benini, "PULPino: A small single-core RISC-V SoC," in *3rd RISCV Workshop*, 2016.

[146] M. Wijtvliet, J. Huisken, L. Waeijen, and H. Corporaal, "Blocks: Redesigning coarse grained reconfigurable architectures for energy efficiency," *International Conference on Field-Programmable Logic and Applications*, pp. 17–23, 2019. DOI: 10.1109/FPL.2019.00013.

[147] A. Quelen, F. Badets, and G. Pillonnet, "A sub-100nW power supply unit embedding untrimmed timing and voltage references for duty-cycled μW-range load in FDSOI 28nm," in *IEEE European Solid State Circuits Conference*, 2017, pp. 279–282. DOI: 10.1109/ESSCIRC.2017.8094580.

[148] V. L. Le and T. T. Kim, "An area and energy efficient ultra-low voltage level shifter with pass transistor and reduced-swing output buffer in 65-nm cmos," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 607–611, May 2018, ISSN: 1549-7747. DOI: 10.1109/TCSII.2018.2820155.

[149] Y. Wang, Q. Sun, H. Luo, X. Wang, R. Zhang, and H. Zhang, "A 48pW, 0.34V, 0.019%/V line sensitivity self-biased subthreshold voltage reference with DIBL effect compensation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 2, pp. 611–621, 2020. DOI: 10.1109/TCSI.2019.2946680.

[150] I. Lee, D. Sylvester, and D. Blaauw, "A subthreshold voltage reference with scalable output voltage for low-power IoT systems," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 5, pp. 1443–1449, 2017. DOI: 10.1109/JSSC.2017.2654326.

[151] *Racyics gmbh*, Apr. 2021. [Online]. Available: `http://www.racyics.de/`.

[152] *Dolphin design*, Apr. 2021. [Online]. Available: `https://www.dolphin-design.fr/`.

Curriculum Vitae

Kamlesh Kumar Singh was born on 05-06-1989 in Bihar, India. After finishing Bachelor's of technology in 2011 at Visvesvaraya National Institute of Technology (VNIT Nagpur) in Maharashtra, India. He studied Master's of technology at Indian Institute of Technology, Bombay (IIT-Bombay) in Mumbai, India. In 2015 he graduated within the Department of Electrical Engineering on Electronic Systems. From 2016 he started a PhD project at Eindhoven University of Technology (TU/e) at Eindhoven, The Netherlands of which the results are presented in this dissertation. Since January 2021 he is employed at Innatera Nanosystems, Delft, The Netherlands.

Acknowledgements

The journey as a PhD student has been a life-changing experience, which shaped me as the person I am today. This journey would have not been possible without the contribution of very special and amazing people around me. Therefore, I would like to thank them here for their support and presence.

First-of-all, I am deeply indebted to my promoter Prof. José Pineda de Gyvez for selecting me for this position. Dear José, I extend my deepest gratitude and appreciation for your supervision, support, motivation, and encouragement during my PhD research period. I am fond of your enthusiasm and the number of ideas given by you during our brainstorming and circuit meetings. You have always been available to advise me and share your immense knowledge with me. Even though you moved to San Jose, USA, you still managed to continuously provide guidance and enough time so that the research continues smoothly. I would also thank you for giving me the valuable opportunity to work at NXP Semiconductors, Eindhoven for over two years. You always inspired me to achieve higher quality, "Work for Glory".

I would like to express my gratitude to my second promotor Prof. Henk Corporaal for his continuous supervision and support. I would like to express my sincere thanks to Dr. Hailong Jiao for his support and guidance. I sincerely appreciate your effort for helping me a lot in the beginning of my PhD project as my daily supervisor and for your continuous help in reviewing and correcting my papers.

Moreover, I would like to use this opportunity to express my appreciation to the defense committee members, Prof. Dr. Mircea Stan from University of Virginia, Prof. Dr. Wim Dehaene from KU Leuven, and Dr. Pieter Harpe from Eindhoven University of Technology for their review and feedback on the thesis and defense. I would like to thank Prof. Dr. Peter Baltus for chairing the defense.

I would like to thank my friend Barry de Bruin for his support to make the BrainWave chip a success, it's been a great pleasure working with you. It would have

been impossible without his architecture and software contributions during the design and testing of the chip. I have learned a lot of exciting things from you. Thanks for being such a nice friend. I wish you all the best.

I am grateful to the Electronic Systems group for providing a great work environment with various activities to keep us motivated. I would like to thank Prof. Twan Basten, Prof. Sander Stuijk, and Marja for their support, kindness, and care. I would also extend thanks to Martijn Koedam for the laboratory support.

My PhD time in TU/e would not have been so amazing without the presence of many amazing colleagues. Having Hadi and Paul in the circuit design discussions accelerated the learning and speed of work. I would like to thank Alessandro, Sayandip, Savas, Paul, Luc, Ilde for making the past five years a fun journey. I hope that we will continue the fun activities in the future. Special thanks to Alessandro for always taking the lead to make important plans for various activities. Special thanks Shima for your kindness, support and the enjoyable time together. I have also enjoyed being together with other people in the group. Thank you, Alejandro, Ali, Alireza, Amr, Berk, Cumhur, Emad, Floran, Gagan, Hamideh, Hossian, Kaniskan, Mahsa, Manil, Marie, Martin, Marzia, Md. Emad, Mojtaba, Mojtaba, Ramon, Rasool, Sajid, Shayan, Wing Tsi, and Zhan. Additionally, I would like to thank Rohit, Swapnali, and Raj for the enjoyable times together. An extended thanks to Rohit for helping me out when I moved to Eindhoven and for awesome food on multiple occasions. I would also like to thank Tathagata dada, Prognya, Mihi, Sumeet, and Peggy for the amazing times we spent together. I would like to thank the friends from Eindhoven cricket club for keeping me fit during the PhD study. I would also like to thank Sumeet, Amir and Rene for including me as a member of Innatera Nanosystems. I would extend my thanks to all the team members of Innatera. I would like to thank Jinbo for accompanying me to Delft and online motivation meetings. Special mention to Aditya, Anushree, Bart, Kasia, Petrut, and Shashanka for the fun times over drinks, lunch and dinner in the office and in Den Haag. Special mention to my friends Akash, Anjali, Ankur, Atishay, Harshad, Mohit, Pandeji, Prashant, Patre, Salil, Umar, Vinayak, and Yogesh. Thank you to all of you.

Last but not least, I would like to express my deepest gratitude to my family for unconditional love, belief, and encouragement throughout the years. Special thanks to Sunita, Jeet, Jahaan, and lovely Jeni for making their home as one of my holiday destination every year. Thanks to my grandparents, father, mother for their blessings.

List of publications

Publications covered in the thesis

- **Kamlesh Singh**, Barry de Bruin, Jos Huisken, Hailong Jiao, Henk Corporaal, and José Pineda de Gyvez, *"Converter-Free Power Delivery Using Voltage Stacking for Near/sub-threshold Operation,"* in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 29, no. 6, pp. 1039-1051, June 2021, doi: 10.1109/TVLSI.2021.3071464.

- **Kamlesh Singh** and José Pineda de Gyvez, *"Twenty Years of Near/Sub-Threshold Design Trends and Enablement,"* in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 1, pp. 5-11, Jan. 2021, doi: 10.1109/TCSII.2020.3040970.

- Barry de Bruin, **Kamlesh Singh**, Jos Huisken, and Henk Corporaal, *"BrainWave: an energy-efficient EEG monitoring system - evaluation and trade-offs,"* ACM/IEEE International Symposium on Low Power Electronics and Design, New York, pp. 181–186, 2020, doi: 10.1145/3370748.3406571.

- **Kamlesh Singh**, Barry de Bruin, Jos Huisken, Hailong Jiao, Henk Corporaal, and José Pineda de Gyvez, *"Voltage Stacked Design of a Microcontroller for Near/Sub-threshold Operation"*, IEEE International System-on-Chip Conference, Singapore, 2019, pp. 370-375, doi: 10.1109/SOCC46988.2019.1570558508.

- **Kamlesh Singh**, Barry de Bruin, Jos Huisken, Hailong Jiao, Henk Corporaal, and José Pineda de Gyvez, *"Voltage Stacking for Near/Sub-threshold Ultra-Low Power Microprocessor Systems"*, IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), San Jose, CA, USA, 2019, pp. 1-2, doi: 10.1109/S3S46989.2019.9320661.

- Wing-Tsi Wong, **Kamlesh Singh**, Jos Huisken, and José Pineda de Gyvez, *"Power and Variation Improved Near-Vt Standard Cell Library for 28-nm FD-SOI"*, IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), San Jose, CA, 2019, pp. 1-2, doi: 10.1109/S3S46989.2019.9320687.
- **Kamlesh Singh**, Hailong Jiao, Jos Huisken, Hamed Fatemi, and José Pineda de Gyvez, *"Low power latch based design with smart retiming"*, IEEE International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, 2018, pp. 329-334, doi: 10.1109/ISQED.2018.8357308.

Publications not covered in the thesis

- **Kamlesh Singh**, O. A. R. Rosas, Hailong Jiao, Jos Huisken, and José Pineda de Gyvez, *"Multi-Bit Pulsed-Latch Based Low Power Synchronous Circuit Design"*, IEEE International Symposium on Circuits and Systems (ISCAS), Florence, 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351251.
- Shima Sedighiani, **Kamlesh Singh**, Jos Huisken, Roel Jordans, Pieter Harpe, and José Pineda de Gyvez, *"An Electromagnetic Energy Harvester and Power Management in 28-nm FDSOI for IoT,"* IEEE Mediterranean Conference on Embedded Computing, 2020, pp. 1-5, doi: 10.1109/MECO49872.2020.9134340.
- Jinbo Zhou, **Kamlesh Singh**, and Jos Huisken, *"Standard Cell based Memory Compiler for Near/Sub-threshold Operation,"* IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4, doi: 10.1109/ICECS49266.2020.9294808.
- Shima Sedighiani, **Kamlesh Singh**, Roel Jordans, Pieter Harpe and José Pineda de Gyvez, *"A Low Power Fully-Digital Multi-Level Voltage Monitor Operating in a Wide Voltage Range for Energy Harvesting IoT,"* IEEE International Symposium on Quality Electronic Design (ISQED), 2021, pp. 13-18, doi: 10.1109/ISQED51717.2021.9424325.
- Barry de Bruin, **Kamlesh Singh**, Ying Wang, Jos Juisken, José Pineda de Gyvez, and Henk Corporaal, *"Multi-level Optimization of an Ultra-Low Power BrainWave System for Non-Convulsive Seizure Detection,"* IEEE Transactions on Biomedical Circuits and Systems (TBioCAS), 2021 (Submitted).