Low-Power Integrated Microcontrollers for Self-Powered Internet-of-Things Applications

by

Daniel S. Truesdell

A dissertation presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

In partial fulfillment of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

May 2021

APPROVAL SHEET

This dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Author: Daniel S. Truesdell

This Dissertation has been read and approved by the examing committee:

Advisor: Benton Calhoun

Committee Member: Steven Bowers

Committee Member: Mircea Stan

Committee Member: N. Scott Barker

Committee Member: Samira Khan

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2021

Abstract

The Internet of Things (IoT) is a maturing technological revolution that is creating new hardware applications in consumer, commercial, and industrial spaces to collect information and generate insights for a variety of benefits. Exponential early growth of the IoT has seen billions of integrated hardware sensing nodes, or “IoT nodes” deployed into action, but this number falls short of early predictions by over an order of magnitude. A major factor preventing the IoT from reaching its originally-projected size is the finite operating lifetime of nodes that are powered by batteries. A node goes offline or “dies” when its battery runs out of charge, and the cost to replace batteries becomes prohibitive as the IoT continues to grow in size. In response to this challenge, there is a growing focus on extending the lifetime of IoT nodes by reducing their power consumption and harvesting ambient energy to reduce or eliminate their dependence on batteries. In the latter case, battery-free or “self-powered” operation can be sustained as long as a node harvests more power than it consumes, and is an attractive solution for enhancing long-term IoT growth. However, without further reducing the power consumption of IoT nodes, the conditions for self-powered operation are currently only met in a fraction of the total applications when high amounts of power can be harvested.

This research makes several contributions to enable self-powered operation in a larger number of IoT applications by investigating integrated circuit design techniques for maximum power reduction. The scope of investigation is on the microcontroller system which forms the data-processing backbone of an IoT node and consists of a digital processor along with its supporting hardware such as memory, clocking circuits, and Input/Output (I/O) peripherals. For digital and memory circuits that are limited by subthreshold leakage currents, a Dynamic Leakage Suppression (DLS) logic style is investigated for low-power operation. A theoretical background and design methodology for Dynamic Leakage Suppression (DLS) logic is derived, and a fabricated test chip is used for experimental verification. Next, a performance-scalable DLS variant is implemented in RISC-V processor and demonstrated in a microcontroller test chip. The application of DLS is extended to memories, and a bitcell

implementation of DLS logic demonstrated in a 16kb Static Random-Access Memory (SRAM) chip. For low-power I/O interfacing, a level converter design based on DLS logic is fabricated and shown to enable low-power signal conversion across a wide voltage range. For clocking circuits, several transistor and architectural-level techniques are explored to reduce power consumption across a wide frequency range. A Hz-range oscillator is designed that leverages tunneling currents through the gate oxide as a low power biasing source. A Frequency-Locked Loop (FLL) design is used for kHz-MHz range operation, and theoretical models are derived for power, energy, and temperature stability. New analog and digital Frequency-Locked Loop (FLL) architectures are designed and fabricated that use frequency multiplication and duty-cycling for power reduction.

Acknowledgments

My time as a graduate student at UVA and the journey toward completing this dissertation have been nothing short of incredible and while I am excited for my next step in life, I am sad that this phase has come to and end. I am first thankful for my advisor, Ben Calhoun, who has supported and inspired me during my journey. Completing a PhD is a transformative experience, at least educationally but often personally as well, and I am fortunate to have had the opportunity to learn from Ben in both categories. He has skillfully provided me with the guidance, resources, flexibility, and opportunities so that I can operate at my own Pareto-optimal point and make the most of my time at UVA. I am so grateful to have had the opportunity to spend years completely immersed in a subject that I enjoy so much with the freedom and support to constantly experiment and learn new things. It is a privilege to be able to find something you are existentially passionate about and truly throw everything you have at it – it is humbling and in many cases surprising to learn what your personal limits are, and also fulfilling to know that you gave your best effort regardless of the outcome. I am thankful for gaining this experience at UVA and knowing that this dissertation is truly a representation of my best effort. Funding from the ASSIST center (NSF NERC ASSIST Center under grant EEC-1160483) is what allowed me to first meet Ben, UVA, and the world of research when I was an undergraduate participating in the REU, and has sustained a significant portion of my research until this day.

There are many other people at UVA who have also contributed to my wonderful experience. I'd like to thank the members of my committee for their time and feedback on this dissertation: Steve Bowers, Scott Barker, Mircea Stan, and Samira Khan. I'm thankful for the opportunities I had to work on outside research projects with Avik Ghosh and Nikhil Shukla, and for the patience and wisdom of Dilip Vasudevan, who had to deal with me when I was an energetic but clueless undergraduate. I'm thankful for Nick Napoli introducing me to supernatural power of Cuban espresso and for the occasions when we were able to escape the Link Lab to grab lunch and chat. Terry Tigner, our research mom, is a miracle worker

who keeps our day-to-day research activities running smoothly. I enjoyed our quick chats and banter whenever I ventured into your office to steal some coffee or pester you to put in another Digikey order. There are many UVA students, past and current, that have impacted my work or experience in some way and will be part of the memory I forged here: Seyi Ayorinde, Farah Yahya, Chris Lukas, Harsh Patel, Arijit Banerjee, Ningxi Liu, Abhiskek Roy, Sheikh Ahmed, Antik Mallick, Nojan Sheybani, Nayiri Krzysztofowicz, Shuo Li, Henry Bishop, Rishika Agarwala, Sumanth Kamineni, Shourya Gupta, Natalie Ownby, Katy Flynn, Xinjian Liu, Jesse Moody, and Pouyan Bassirian. Many late nights testing in Rice Hall, trips to conferences (ISSCC 2020 was the best ISSCC), and Afghan Kabob dinners in the Link Lab cafeteria during tapeouts were had with these great people. I have been lucky to work with Jacob Breiholz on much of the work in this dissertation. Our skill sets have been nearly perfectly complementary to tackle the challenges of DLS logic and many of the ideas and concepts that we created would never have come to fruition without your knowledge and experience in digital design. The results of our collaboration guide my personal belief that you can learn more and achieve greater success by sharing your knowledge and working together with others, rather than trying to do everything yourself to get all the credit. I have also appreciated the friendship of Anjana Dissanayake, who is an amazingly talented designer and always has something interesting to say. Several of my circuit designs would not function without your contributions to the 'Danplifier', and I hope there will be more chances in the future for interesting technical discussions and philosophical debates.

Outside of UVA, I have many other people in my life that I'd like to acknowledge and thank. My brother (who is now a fellow UCF knight!), grandparents, and relatives who have always been very supportive. Cheers to my friend Nick who built many home science and engineering projects with me through grade school and is now also completing his PhD, and to some of my other friends: Noble, Dorris, Duncan, Tyler, Squared, and Evan – our many nights of gaming have been probably the single biggest source of stress relief during my entire time in grad school - thank you for keeping me grounded and sane. I'd like to officially endorse Café Bustelo as the greatest coffee in the world which has given me the energy to complete most of the work in this dissertation. I have had at least one cup of it per day on overage over the last 5 years.

To my mom and dad, you are a constant source of inspiration for everything I do in life. I wouldn't have the motivation to chase my dreams or the means to succeed in doing so without the constant love and support you have given me. I am grateful that you stimulated my interest in science and engineering from such a young age and did everything possible to encourage me to learn, experiment, and follow my interests as long as I am happy. I will always be grateful for (and try as hard as I can to perpetuate one day as a parent) everything you have done to provide me with this opportunity. This dissertation is just as much a representation of your time and effort as it is my own.

Finally, I'd also like to thank Sierra for her love and and support over the last (nearly) 6 years. It was a big change in our lives to uproot ourselves and move to Virginia, and you did it without batting an eye – although in hindsight it was an easy decision cause living in VA beats the heck out of living in FL. You have supported me in so many ways while I have been in school, and I will always be grateful that you have stood by me through years of late nights and long weekends. I'm also so happy and proud that you were able to join me at UVA for a short time to chase your own dream, graduate with your Master's degree, and land a perfect job. Now we are both UVA alumni!

List of Figures

1.1 Example block diagram of a modern IoT system with integrated wireless circuits, sensors, energy harvesting and power management circuits, and a microcontroller. 5
1.2 Energy harvesting and power management architecture of an IoT node. . . . 6
1.3 Transient operating conditions for an IoT node that waits in a standby mode until enough energy has been harvested and stored that it can enable high-power circuits to obtain, process, and transmit a data sample. 9

2.1 Power consumption versus operating frequency (f_{clk}) of a 32-bit ripple-carry adder in 65nm, built with traditional static-Complementary Metal-Oxide Semiconductor (CMOS) logic gates, while operating at a 0.5V supply voltage V_{dd}. 12
2.2 Traditional static-CMOS implementation of a Boolean inverter with illustration of leakage currents that travel through the NMOS device. 13
2.3 Breakdown of drain current in an NMOS transistor as it would be used in a digital logic gate. Total drain current versus gate voltage is shown for supply voltages of 1.2V and 0.5V. At negative gate voltages, either Gate-Induced Drain Leakage (GIDL) or gate leakage will cause an increase in total current, depending on the supply voltage. 16
2.4 Illustration of the stack effect for reducing subthreshold leakage and its application in power gating to cut off subthreshold leakage current to a digital circuit. 17
2.5 (a) Traditional static-CMOS inverter, (b) Dynamic Leakage Suppression (DLS) inverter, (c) Illustration of leakge reduction behavior in a DLS inverter. 20
2.6 Pull-down network of a DLS inverter with leakage current sources that affect the DC value of V_y. At low V_{dd}, contention between subthreshold currents determines V_y, while at high V_{dd}, it is contention between gate and junction leakage that determines V_y. Models for both cases agree well with simulated data. 22
2.7 Schematic of a DLS inverter and its simulated DC transfer characteristic, showing the output voltage Y as well as the internal nodes V_y and V_x. 25
2.8 Derivation of input threshold voltage levels and output voltage levels of the DLS gate, and illustration of how the noise margins (NM) and switching margins (SM) are derived from these values. 27
2.9 Equivalent schematics for determining the input threshold voltages for a DLS logic gate. Part (a) shows the pull-down network to solve for V_{il}, part (b) shows the pull-up network to solve for V_{ih}. 29
2.10 Modeled and simulated input threshold voltages (V_{il} and V_{ih}) of the DLS inverter versus V_{dd}. Low-V_{dd} models are $V_{il,1}$ and $V_{ih,1}$ from (23) and (21), and the high-V_{dd} model for V_{ih} uses $V_{ih,2}$ from (25). 31
2.11 Sources of drive current and parasitic leakages currents that determine the maximum output voltage V_{oh} (a) and minimum output voltage V_{ol} of the DLS gate. 32

2.12 Transient simulation of a DLS inverter switching from output low to output high at V_{dd}=0.5V. The operating regions of each transistor are illustrated at different key points throughout the transition. . . . 34
2.13 Modeled and simulated gate delay (output low to high) of a DLS inverter versus supply voltage V_{dd}. . . . 35
2.14 Modeled and simulated leakage current components in a DLS gate with input A=0V versus supply voltage V_{dd}. Models are shown by lines, simulated values are shown by the marker dots. . . . 36
2.15 Modeled and simulated leakage power P_{leak} for two different DLS gate designs with the modeled optimum V_{dd} for minimizing leakage power. . . . 38
2.16 Fabricated DLS test chip in 65-nm CMOS. . . . 39
2.17 Measured Voltage Transfer Curves (VTCs) from 32 LVT-LVT test DLS inverters in a single die operating at V_{dd}=0.25V. Input and output voltages are both normalized to the V_{dd}. . . . 41
2.18 Measured inverter delay and inverter leakage versus supply voltage for several DLS flavors, measured across 20 dies. Leakage for a traditional High Threshold Voltage (HVT) static-CMOS inverter is also shown. Bolded lines show the average values across all measured dies. . . . 42
2.19 Circuit setup for measurement of DLS inverter VTCs (a) and functional diagram of the implemented Multiply-Accumulate (MAC) circuit. . . . 43
2.20 Measured VTCs from 32 LVT-LVT test DLS inverters in a single die operating at V_{dd}=0.25V. Input and output voltages are both normalized to the V_{dd}. . . . 43
2.21 Measured DLS inverter DC input and output characteristics versus supply voltage, obtained from all 32 measured VTCs for each of the DLS flavors in one die. . . . 44
2.22 Measured DLS inverter operating margins versus supply voltage, obtained from all 32 measured VTCs for each of the DLS flavors in one die. . . . 45
2.23 Measured distribution of input threshold voltages (V_{il} and V_{ih}) and output voltage levels (V_{ol} and V_{oh}) of all DLS structures and all chips at a V_{dd} of 0.25V. With 32 measured VTCs from each of 20 chips, this yields 640 measured points for each structure type. . . . 46
2.24 Layout implementation of a DLS inverter (a) and standard cell array implementation for Forward Body Bias (FBB) IO-IO DLS logic flavor (b). . . . 49
2.25 Measurements of the DLS Multiply-Accumulate (MAC) circuits with comparison to HVT CMOS. The maximum operating frequency versus supply voltage is shown, as well as a pareto-optimal plot for power versus energy. . . . 50
2.26 Measured leakage power versus supply voltage of the DLS and static-CMOS MAC circuits. . . . 51
2.27 Design space for low-power and low-frequency microprocessors. DLS-based works are limited by their maximum speed [1,2], while traditional CMOS-based works are limited by their leakage currents [3]. . . . 53
2.28 (a) Design of Scalable Dynamic Leakage Suppression (SDLS) logic gate, (b) DLS-based operation of Scalable Dynamic Leakage Suppression (SDLS) gate when the control voltage is low, and (c) CMOS-based operation of SDLS gate when control voltage is high. . . . 54

2.29 Illustration of how the control voltage affects the operating parameters of the devices in the SDLS gate. Scaling the control voltage up (higher V_c means higher V_{nc} and lower V_{pc}) gives the gate more drive current via M_{nc} and M_{pc}, but also increases their leakage current as well. 55
2.30 Schematic diagram of Voltage Scaling Controller (VSC) and measured oscilloscope plot of the VSC outputs while the control value *Vc SEL* is swept from its minimum to its maximum value. $V_{analog} = 1.2V$, $V_{div} = 1.0V$. 56
2.31 Design of SDLS inverter to be used as a clock delay cell for critical path replication. Changing the widths of M_{nc} and M_{pc} allows the clock cell to achieve a different delay scaling across V_c that accommodates the different intrinsic delay characteristics of SDLS standard cells. 57
2.32 Schematic diagram of the SDLS Adaptive Clock Generator (ACG). A 3-stage SDLS-based ring oscillators is divided down in frequency by a 2-phase programmable frequency divider that allows for customizable frequency and duty-cycle. 58
2.33 Measured frequency and power of the Adaptive Clock Generator (ACG). Frequencies of each individual SDLS Ring Oscillator (RO) are recorded, and the frequency tuning resolution is measured by adjusting the frequency divider value. 59
2.34 Architecture of the SDLS-based RISC-V microprocessor. The core domain includes the RISC-V core, while the uncore domain includes interconnects, registers, and peripherals. A power and timing management subsystem includes a programmable reference current generator, the adaptive clock generator, and the voltage scaling controller 60
2.35 Annotated micrograph of the SDLS microprocessor fabricated in 65-nm CMOS. 61
2.36 Oscilloscope measurement of runtime Dynamic Voltage and Frequency Scaling (DVFS) with core dithering between two V_c *SEL* modes while computing the Fibonacci sequence. General-Purpose Input/Output (GPIO) output bits indicate when a new Fibonacci number is computed and when the total sequence (numbers 1-46) are correct. 61
2.37 Performance of the SDLS core and uncore at 0.6V V_{dd} across all possible DVFS modes, obtained by sweeping V_c *SEL* across all possible values. 62
2.38 Performance scaling ranges (core only) achievable through DVFS (sweeping V_c *SEL*) at each supply voltage point. 63

3.1 Traditional 6T SRAM design (a), and implementation of 6T bitcell using DLS inverters (b). 65
3.2 Concept of 6T SRAM bitcell using DLS inverters (b). 68
3.3 Potential for leakage currents through the Bit Line (BL) to disrupt the stored data in adjacent DLS bitcells. 69
3.4 Implementation Word Line (WL) overdrive to suppress leakage current through the access transistors. 70
3.5 Monte Carlo simulation of the leakage reduction through the access transistors for the traditional NMOS access transistors and the utilized PMOS access transistors with overdrive WL voltage. 71
3.6 Implementation of DLS SRAM array using DLS level shifter for WL boosting. . 72
3.7 Monte Carlo simulation of the Data Retention Voltage (DRV) of a traditional 6T static-CMOS bitcell and the proposed DLS bitcell, showing an improvement in minimum retention V_{dd} from the DLS bitcell. 73

3.8 Schematic of a traditional static-CMOS digital buffer (a) and schematic of DLS-based level converter (b). 74
3.9 Schematic of DLS-based inverter used in the level converter design and simulated DC transfer characteristics at V_{ddh}=1.2V with comparison to a traditional static-CMOS digital inverter with skewed Pull-Up (PU) and Pull-Down (PD) sizes. 75
3.10 Modeled and simulated input logic-high threshold voltage V_{ih} for the DLS level converter versus the sizing ratio its transistors, and the susceptibility of V_{ih} to process corner versus the nominal (TT) case. 76
3.11 Architecture of the 16Kilobit (kb) DLS SRAM macro. 77
3.12 Fabricated 16kb DLS SRAM test chip in 65-nm CMOS. 78
3.13 Measured leakage current of the DLS bitcell and leakage power of the full SRAM macro across V_{dd}. Reference [a] corresponds to [3], [b] corresponds to [4], and [c] corresponds to [5]. 78
3.14 Measured operating frequencies of the DLS SRAM (both reading and writing) and read energy (energy per access per bit) versus V_{dd}. 79
3.15 Measured leakage power and read access frequency of the DLS SRAM versus V_{dd}, shown across a temperature range of 0°C to 60°C. 80

4.1 Classification of common designs for clocking circuits. 81
4.2 Classic Relaxation Oscillator (RXO) architecture. 83
4.3 (a) Phase-Locked Loop (PLL) architecture for frequency synthesis, (b) Frequency-Locked Loop (FLL) architecture for frequency synthesis, and (c) Frequency-Locked Loop for frequency generation. 86
4.4 (a) Design of Frequency-to-Voltage Converter (FVC), (b) operating waveforms of the Frequency-to-Voltage Converter (FVC), and (c) operating waveforms for the FLL. 88
4.5 Layout diagram of implementation for gate leakage-based resistive segment. . 92
4.6 Comparison between resistive densities of traditional on-chip polysilicon resistors and the effective resistance of the transistor gate. To create a 1GΩ effective resistance, the two structures differ in their two-dimensional silicon area by over 10^{15}x. 93
4.7 Traditional technique for trimming an on-chip resistor (a), design goal for resistive segment bypass switches (b), implementation with standard resistors (c), and implementation with proposed gate leakage-based resistive segments (d). 94
4.8 Proposed reference current generator based on gate leakage that enables a tunable transistor width to scale the value of the reference current. 95
4.9 Parallel arrangement of individual gate leakage segments to enable tunable size adjustment (a) and concept applied to traditional resistances by considering the gate leakage as a conductance G. 96
4.10 Gate leakage-powered oscillator design with tunable I_{ref} generator and dual-phase operation. 97
4.11 Fabricated chip in 65-nm CMOS that contains the proposed gate-leakage powered Relaxation Oscillator (RXO) and an on-chip temperature sensor. . . 99
4.12 Measured frequency of the gate leakage-powered RXO across temperature, shown for different structures that use different device types. 99

4.13 Measurements of dynamic temperature compensation for the gate leakage-powered RXO showing the readout from the on-chip temperature sensor, the dynamically-updated I_{ref} tuning code, and the compensated and uncompensated RXO frequencies. 100
4.14 Allan deviation measurement for the gate leakage-powered RXO showing a sub-1000ppm deviation across averaging times and a floor of 300ppm. 101
4.15 Design of the analog FLL (a) and implementation of R_{ref} to enable passive temperature compensation (b) 103
4.16 Schematics of the analog FLL loop amplifier (A) and Voltage-Controlled Oscillator (VCO) (b). 104
4.17 Modeled and simulated energy consumption of the analog FLL, showing breakdown of energy between the divider, VCO, and I_{ref} Models used are from (70). 107
4.18 Annotated micrograph of the fabricated analog FLL design in 65nm CMOS. . . 108
4.19 Measured start-up response of the FLL, showing a 10ms settling time. Dashed line shows visualization of FLL amplifier settling without the added modulation from the I_{ref} startup. 108
4.20 Measured power and energy consumption of the FLL versus the output frequency, obtained by changing the divider value from N=2 to N=16 (a) and the Temperature Coefficient (TC) of each output frequency measured from -20°C to 60°C (b). 109
4.21 Measured temperature compensation ability of R_{ref} shown by measuring F_{out} over the full temperature range for each discrete R_{ref} trim setting (a) and sensitivity of F_{out} to the input reference voltage V_{in}. 110
4.22 Sensitivity of the FLL frequency and temperature stability to supply voltage, measured from the nominal 0.6V supply up to a 0.8V supply. Higher supply voltages increase the output frequency and reduce its temperature stability (higher TC). This effect impacts the FLL equally regardless of what output frequency (divider N) it is running at. 111
4.23 Allan deviation measurement of the analog FLL showing a flicker noise-limited floor of 300pm at an averaging time of 100ms. 112
4.24 Generalized architecture of a digital FLL. 114
4.25 Simplified schematic and timing diagram of the FVC (a), illustration of locking probability during a decreasing frequency search (b), and resulting voltage/frequency locking range (c). 117
4.26 Independent contributions of the amplifier, comparator, reference resistor, and locking circuitry to the total frequency inaccuracy of the Digital Frequency-Locked Loop (DFLL), shown as a function of the TC of each component. A temperature range of 100°C (ΔT=50) is used. 119
4.27 Architecture and timing waveforms of the proposed duty-cycled DFLL. 120
4.28 DFLL energy consumption during active and duty-cycled modes (a). Digitally-controlled Oscillator (DCO) energy is the same for both active and duty-cycled modes, and the divider energy is only present in the active mode. (b) shows total energy reduction ((85) divided by (74)) versus duty cycle ratio, shown for different divider values. 122
4.29 Full schematic of the duty-cycled DFLL including I_{REF} generation circuit and always-on digital timing generation block. 123

4.30 Schematic of (a) the V-I converter with self-biased amplifier and (b) the clocked comparator based on a StrongArm latch with tunable offset. 124
4.31 Timing diagram of the DFLL booting up, locking, duty-cycling, and waking up to re-lock. Diagram of locking and duty-cycling algorithm shown on right, where processes with dashed borders indicate that they only proceed after waiting for a timer. 125
4.32 Modeled maximum open-loop frequency drift versus duty-cycle interval shown for temperature fluctuations of 0.1°C/s, 1°C/s, and 10°C/s (a), and average energy per cycle of the DFLL required (based on duty-cycle rate) to limit open-loop frequency drift to within ΔF_{lock} versus the rate of ambient temperature drift (b). 128
4.33 Annotated chip micrograph of the proposed DFLL in 65-nm CMOS. 130
4.34 Measured and power and energy consumption versus output frequency F_{out} measured at 20°C. 130
4.35 Measured output frequency F_{out} versus temperature, shown for divider values N=2, N=5, and N=10 on four dies (a) from 0°C to 100°C. TCs of the four measured dies are shown across the full range of output frequencies (N=1 to N=10) (b). 131
4.36 Measured transient operation of the DFLL during a positive and negative 20°C step in temperature, shown for duty cycle rates of t_{DC}=0.1s (r_{DC}=0.36), t_{DC}=1s (r_{DC}=0.06), and t_{DC}=8s (r_{DC}=0.01). For reference, the open-loop DCO response (with temp. change) and baseline closed-loop operation (t_{DC}=1s without temp. change) are shown. 132
4.37 Allan deviation measurement of the DFLL at F_{out}=560kHz (N=10) in an indoor environment at room temperature (20°C), shown for open-loop operation as well as duty-cycle intervals of 1s and 14s. 133
4.38 Measured supply voltage sensitivity for ±50mV of deviation from the nominal supply voltage (a), and reference voltage sensitivity for ±20mV of deviation from the nominal reference voltage. Both measurements taken at 20°C. 133
4.39 Power versus frequency design space for on-chip clocks. Triangle markers show FLL designs, square markers show RXO designs. For very low-power designs, labels are printed to indicate the TC of the design. 135
4.40 Measured supply voltage sensitivity for ±50mV of deviation from the nominal supply voltage (a), and reference voltage sensitivity for ±20mV of deviation from the nominal reference voltage. Both measurements taken at 20°C. 136

List of Tables

1 Sources of ambient energy and the amount of power that can be harvested from them. Table and data are adapted from [6] 7

2 Leakage current sources that affect node V_y. 23
3 Definitions of the DC characteristics of the DLS logic gate. 28
4 Summary of DLS flavors in a 65-nm technology that intend to reach different reliability and performance points. 40
5 Measurement of variation in the DC input and output characteristics of the DLS inverters from 20 chips operating at a 0.25V V_{dd}. 47
6 Measurement of variation and yield in the operating margins (SM and NM) of the DLS inverters from 20 chips operating at a 0.25V V_{dd}. 48
7 Performance summary of SDLS RISC-V system design (includes core, uncore, VSC, and ACG) with comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (09/18/2019), and more recent works are not shown here. Reference publication for this work is [DT17]. 64

8 Performance summary of DLS SRAM and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (6/16/2020), and more recent works are not shown here. Reference publication for this work is [DT7]. 80

9 Measurement summary of the gate leakage-powered RXO based on different compensation and tuning approaches. 101
10 Performance summary of analog FLL design and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (10/10/19), and more recent works are not shown here. Reference publication for this work is [DT15]. 113
11 Performance summary of digital FLL design and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (12/28/2020), and more recent works are not shown here. Reference publication for this work is [DT2]. 134

Contents

Abstract iii

Acknowledgments v

List of Figures viii

List of Tables xiv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Thesis and Contributions . . . 2
  - 1.2.1 Digital Logic . . . 3
  - 1.2.2 Memory . . . 4
  - 1.2.3 Clocking Circuits . . . 4
- 1.3 Self-Powered IoT Systems . . . 5

2 Digital Logic 11
- 2.1 Introduction . . . 11
  - 2.1.1 Leakage Current . . . 13
  - 2.1.2 Leakage Reduction for Digital Circuits . . . 17
  - 2.1.3 DLS Logic . . . 19
- 2.2 Modeling for DLS Logic . . . 22
  - 2.2.1 DC Switching Characteristics . . . 22
  - 2.2.2 Transient Characteristics . . . 33
  - 2.2.3 Power . . . 36
  - 2.2.4 Measurement Results . . . 39
  - 2.2.5 Conclusion . . . 51
- 2.3 Performance Scaling for DLS Logic . . . 53
  - 2.3.1 SDLS Design . . . 54
  - 2.3.2 SDLS Supporting Hardware . . . 56
  - 2.3.3 SDLS RISC-V Core . . . 59
  - 2.3.4 Measurements . . . 61
  - 2.3.5 Conclusion . . . 64

3 Memory 65
- 3.1 Introduction . . . 65
  - 3.1.1 SRAM power reduction . . . 67
- 3.2 DLS SRAM Design . . . 68
  - 3.2.1 DLS Bitcell . . . 68
  - 3.2.2 DLS level converter . . . 73
  - 3.2.3 Measurement Results . . . 77
  - 3.2.4 Conclusion . . . 80

4 Clocking Circuits **81**
4.1 Introduction . . . 81
4.1.1 The Relaxation Oscillator . . . 83
4.1.2 The Frequency-Locked Loop . . . 85
4.2 Hz-Range Relaxation Oscillator . . . 91
4.2.1 Gate Leakage as a Reference Resistance . . . 91
4.2.2 Gate Leakage-Powered RXO Design . . . 95
4.2.3 Measurement Results . . . 98
4.2.4 Conclusion . . . 102
4.3 Analog FLL . . . 103
4.3.1 Design and Analysis . . . 103
4.3.2 Measurement Results . . . 107
4.3.3 Conclusion . . . 113
4.4 Duty-Cycled Digital FLL . . . 114
4.4.1 Digital FLL Design . . . 114
4.4.2 Steady-State Frequency Inaccuracy . . . 116
4.4.3 Duty-Cycling Concept . . . 120
4.4.4 Implementation . . . 122
4.4.5 Transient Frequency Inaccuracy . . . 129
4.4.6 Measurement Results . . . 129
4.4.7 Conclusion . . . 134
4.5 Design Space Summary . . . 135

5 Conclusion **137**
5.1 Summary of Results and Contributions . . . 137
5.1.1 Digital Logic . . . 137
5.1.2 Memory . . . 139
5.1.3 Clocking Circuits . . . 139
5.2 Future Work . . . 141

Appendices **xvii**

A Published Works **xviii**

B Abbreviations **xxi**

C Algorithms **xxiv**

D References **xxv**

1 Introduction

1.1 Motivation

The Internet of Things (IoT) has grown significantly over the last decade, fueled in part by a self-fulfilling prophecy that it will unlock a massive technological revolution. Although this revolution is still underway, the success of the IoT so far is unquestionable: In 2018, the IoT industry was valued at over 200 billion U.S dollars, which is expected to grow to over 1300 billion U.S. dollars by 2026 [7]. A full-scale IoT means that billions of hardware sensor nodes, or "things", will be deployed in consumer, commercial, and industrial spaces to facilitate the collection and exchange of information for generating valuable insights and feedback. Further growth of the IoT depends on an increase in the number of "things", which places demand on sensor systems to support more applications, increased connectivity, and higher reliability. One important challenge in IoT scaling is caused by the limited lifetime of hardware sensor nodes that are powered by batteries. A sensor goes offline, or "dies", when its battery runs out of charge, and the cost to replace batteries becomes prohibitive for a fully-scaled IoT [8]. In response to this threat, IoT nodes have focused on both reducing their power consumption as well as harvesting ambient energy, usually from solar or thermal sources, to reduce or eliminate their dependence on batteries.

The exact hardware implementations of IoT nodes can vary, with some factors being heavily application-dependent such as wireless connectivity, standards compliance, sensing modalities, and energy harvesting modalities. Common to nearly every sensor node, however, is the microcontroller system, which consists of a digital processor, memory, a clock source, and communication periphery such as Serial-Peripheral Interface (SPI), I2C, and GPIO interfaces. The microcontroller forms the data processing backbone of a sensor node, and has proliferated alongside the IoT industry with sales currently at tens of billions of U.S. dollars with significant projected growth over the next several years [9, 10]. Commercial microcontrollers for the IoT have adapted to reduce active power to the μW-level and offer flexible sleep- or standby-modes that enable compatibility with mature harvesting technologies,

such as solar and thermal, that provide roughly 10μW–10mW of power in a limited number of applications where these sources are available. Further reduction in microcontroller power has been fundamentally limited by the subthreshold leakage of the transistors used in CMOS circuits, such as the logic gates in the digital processor or the bit cells in the memory (SRAM). Other barriers include the dynamic power and DC biasing used in the clocking circuits. Continuous innovation in energy harvesting technology has led to a number new harvesting modalities that offer less power than existing sources but could unlock many new applications. For example, ambient Radio Frequency (RF) harvesting from ambient signals such as WiFi typically provides nW-levels of power [11, 12]. Similarly, nW of power can be harvested from ambient humidity using nanometer-scale protein wires [13]. Biological sources and bio-fuels can provide a wide array of power from mW to pW [14–16], and recently, the thermal motion of free-standing graphene has been shown to provide pW- to nW-levels of power [17]. To leverage emerging energy-harvesting sources and grow the number of applications accessible by the IoT, hardware nodes must continue to break barriers in reducing their power consumption so that they can sustain operation at the nW-level.

This dissertation focuses on reducing the power of microcontroller systems below the existing power floor to enable IoT growth in new applications under emerging energy harvesting modalities. Several fundamental design techniques on the circuit and architectural levels, such as logic gate design, DC biasing, and architectural duty-cycling, are demonstrated that overcome existing performance limitations and can be applied to clocking, processing, memory, and I/O subsystems of the microcontroller. The proposed techniques leverage unconventional transistor configurations but are fully compatible with commercially-available IC fabrication abilities. The efficacy of these techniques in reaching state-of-the-art pW-nW power levels is demonstrated on component and system-levels through several proof-of-concept test chips.

1.2 Thesis and Contributions

Self-powered operation extends the lifetime of IoT nodes to support increased growth and sustainability of the IoT. The ability for self-powered operation depends on both the amount of energy that can be harvested from the environment as well as the power consumption of the node itself. Decreasing the power consumption of a node will improve the ability for

self-powered operation and therefore create new applications, but will also result in degraded performance. In some cases, an extreme tradeoff between power and performance is justified to allow self-powered operation in select applications where ambient energy is very scarce. This dissertation demonstrates design techniques applicable to digital logic, memory, and clocking circuits that collectively advance the state-of-the-art for low power design and enable larger performance trade-offs in microcontrollers for self-powered applications that prioritize extremely low-power operation.

1.2.1 Digital Logic

Total power in digital circuits depends mainly on the sum of dynamic power and leakage power. In static CMOS digital circuits, dynamic power is proportional to the operating frequency, so operating the circuit slowly can yield a very small amount of dynamic power, however leakage currents are always present and currently bottleneck power reduction efforts. This obstacle is hard to avoid in static CMOS circuits due to the way that transistors are arranged in each logic gate. This research leverages a new logic gate design named Dynamic Leakage Suppression (DLS) logic that reduces subthreshold leakage currents by using a new transistor configuration. A theoretical background on the operation of DLS logic is derived to show the factors that influence leakage and delay, and a design exploration is performed to analyze the design approach, potential performance knobs, and performance limits for DLS logic. These results help pave a path toward implementing new DLS libraries in other technologies or to meet different performance requirements. Additionally, test chips in 65nm are fabricated to show two new DLS implementations: First, a performance-scalable DLS-based logic gate design is used to implement RISC-V digital processor that enables a smooth performance tradeoff between DLS logic and static-CMOS logic. Second, a forward body-biased DLS design is used to implement a MAC circuit, showing a proof-of-concept technique for improving the reliability and performance of DLS logic. Together, these results advance the state-of-the-art for low-power digital circuit design.

1.2.2 Memory

The power consumption of a memory array is similar to a digital circuit operating at a very low frequency. Typical memory designs contain two cross-coupled inverters per bitcell which adds up to tens of thousands of inverters, and only a tiny fraction are switched per clock cycle on average, so the total power of memory circuits is easily dominated by subthreshold leakage from the bitcells. For true power reduction of the microcontroller system, the power of the memory must be reduced along with power of the digital logic. A new Static Random-Access Memory (SRAM) bitcell is designed that uses cross-coupled DLS inverters to significantly the subthreshold leakage. The design and performance of the DLS inverter is reanalyzed from this context to ensure reliable operation. A full SRAM macro implementation is developed for use with the DLS bitcell, and is fabricated in a 65nm test chip.

1.2.3 Clocking Circuits

Clocking circuits can be dominated by either dynamic or leakage power depending on the operating frequency and architecture. At high output frequencies, dynamic power from active circuits tends to dominate, while at low output frequencies the biasing and leakage currents can dominate. For low frequency generation, a traditional Relaxation Oscillator (RXO) architecture is shown that uses an ultra-low magnitude bias current derived from tunneling currents through the gate oxide of a transistor. The same technique can also be used to build a low-power temperature sensor that can be used to dynamically tune the biasing current of the RXO to compensate its performance across temperature. For high frequency generation, new modifications to the Frequency-Locked Loop (FLL) architecture are introduced to improve energy efficiency by reducing dynamic power. An analog FLL design is developed that uses a frequency divider to boost the output frequency and reduce the relative contribution of static energy. To reduce the dynamic power of the newly added frequency divider, a digital FLL design is created that duty-cycles the divider for short intervals without causing the frequency to become unlocked. The RXO, analog FLL, and digital FLL are all demonstrated in 65nm test chips.

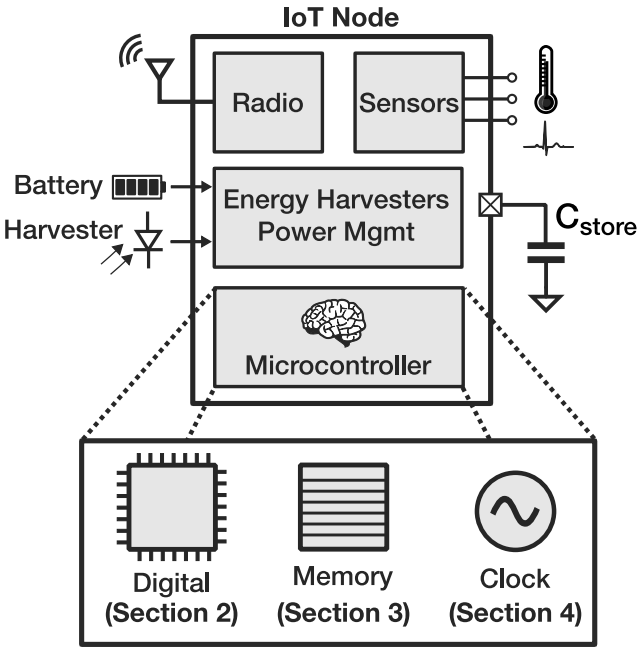

Fig. 1.1. Example block diagram of a modern IoT system with integrated wireless circuits, sensors, energy harvesting and power management circuits, and a microcontroller.

1.3 Self-Powered IoT Systems

Fig. 1.1 shows a block diagram of a traditional IoT node, which consists of circuits for wireless communication (radio), sensing circuits such as analog-to-digital converts, energy harvesters and power management circuits, and the microcontroller, which is the focus of this dissertation. A commercially-available example of such a system is the ESP32-PICO-D4, from Espressif Systems [18], which is a System-in-Package (SIP) that contains a 2.4GHz WiFi and Bluetooth Low-Energy (BLE) module, on-chip and discrete (crystal oscillator) clocks, and the ESP32 microcontroller fabricated in 40nm CMOS that contains a Dual-core 32-bit processor with several hundred Kilobytes (kBs) of on-chip static random-access memory (SRAM) to hold data and instructions. This type of node contains most if not all of the critical hardware for many sensing applications in the IoT. For example, an environmental sensor could be implemented that samples ambient temperature, humidity, and light intensity, and transmits the data back to a base station when a request is made or automatically at some semi-regular interval. IoT applications that require any form of physical actuation (for example, in the form of Light-Emitting Diodes (LEDs) or motors) generally consume too much power to

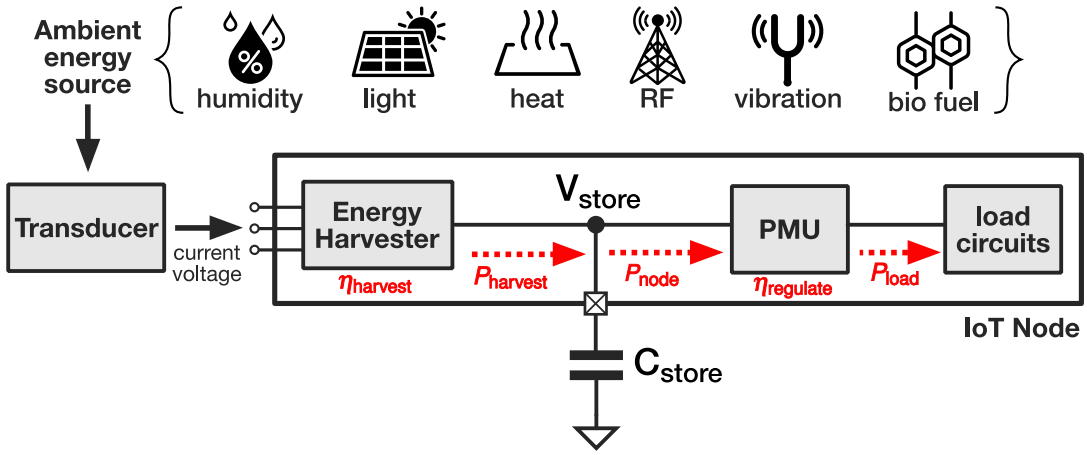

Fig. 1.2. Energy harvesting and power management architecture of an IoT node.

be compatible with self-powered operation (batteries are almost certainly required), so their associated hardware nodes are precluded from the scope of this dissertation. In general, this node will operate from some stored and finite quantity of energy E_{store}, and if we temporarily take a simplified approach to assume that the node will consume that energy at a constant rate P_{node}, then the lifetime of the node follows the relationship

$$\text{Lifetime} \propto \frac{E_{\text{store}}}{P_{\text{node}}} \tag{1}$$

Naturally, using a larger battery to provide more E_{store} or making a more energy-efficient node to decrease P_{node} will increase the lifetime of the overall system. Energy harvesting is an increasingly popular method to effectively increase E_{store} by harvesting freely-available ambient energy to supplement the original finite amount of stored energy. Fig. 1.2 shows an example energy harvesting and power management architecture for an IoT node, where a transducer is used to convert some form of ambient energy into electrical signals (current and voltage) which can be processed (rectified and regulated) by a dedicated energy harvesting circuit. There are many types of energy that can be harvested which are summarized by Table 1. The harvester circuit consumes some amount of quiescent current, so its efficiency η_{harvest} is a design concern to ensure maximal harvested power P_{harvest}. This harvested power can be used to recharge a battery, however the nature of batteries to store energy chemically leads to decreased endurance (cycle life) that causes their maximum capacity to

Energy Source	Harvested Power		Reference
Vibration			
Human movement	1 – 10	μW	[19, 20]
Machine vibration	10 – 100	μW	[19, 21]
Heat			
Human body heat	9 – 25	μW	[19, 22]
Machine heat	1 – 10	mW	[19]
Environmental			
Moonlight	1.5	nW (0.54mm^2)	[23]
Indoor (bright office)	10	μW/cm^2	[19]
Outdoor (direct sunlight)	10	mW/cm^2	[19]
Atmospheric moisture	50-400	nW (1mm^2)	[13]
Radio Frequency			
GSM900	36	nW/cm^2	[11, 12]
GSM1800	84	nW/cm^2	[11, 12]
3G	12	nW/cm^2	[11, 12]
WiFi	180	pW/cm^2	[11, 12]
Biological			
Biochemical (sweat)	1.2	mW/cm^2	[15]
Biochemical (glucose/O_2)	2.2	nW/cm^2	[16]
Biological endocochlear potential	1.1-6.3	nW (99mm^2)	[14]

Table 1. Sources of ambient energy and the amount of power that can be harvested from them. Table and data are adapted from [6]

deteriorate the more times that it is discharged and recharged [24]. Moreover, batteries suffer from self-discharge that would cause a fully-charged battery to eventually lose its stored energy over an extended period of time even if nothing is attached to it, which reduces or eliminates the benefit of reducing circuit power consumption. Super-capacitors (C_{store}) offer a more robust means for storing energy that behave more ideally over time and are therefore a common energy storage mechanism for long-lifetime IoT nodes. The maximum amount of energy that can be stored on C_{store} is

$$E_{\text{max}} = \frac{1}{2} C_{\text{store}} V_{\text{max}}^2 \tag{2}$$

Where V_{max} is the maximum allowable voltage on C_{store}. Intuitively, a larger capacitance is desirable, and it is desirable to set V_{max} to be as high as is reasonable, typically around 3.3V to 5V, to maximize the potential amount of stored energy. Of course, the load circuits of the node (such as the microcontroller) do not need to run from such a high voltage, so an on-chip Power Management Unit (PMU) with efficiency η_{regulate} lowers the voltage (for example, with a linear or switched-capacitor regulator) to a value usable by the remainder of the

circuits in the node. The resulting behavior of the system is a push-pull relationship between P_{harvest} and P_{node}, where the stored voltage V_{store} (and therefore E_{store}) will increase if $P_{\text{harvest}} > P_{\text{node}}$ and decrease if $P_{\text{harvest}} < P_{\text{node}}$. Depending on the operating requirements for both the regulator and the load circuits, V_{store} must not decrease below a certain value V_{kill}, otherwise the node will fail to operate and potentially lose its data. Modern IoT System-on-Chips (SoCs) can include integrated Non-Volatile Memorys (NVMs) can be used to store critical instructions and data in the event of power loss, and energy harvesters are designed for the ability to cold-boot even if there is no stored energy in the system [25], but node death is still generally undesirable due to both the interruption to operation and also the large energy overhead associated with resuming system operation (reading from the NVM and re-starting the energy harvester). Therefore, for fully self-powered and "always-on" operation, the first condition to be met is simply that over an infinite period of time, more power must be harvested than that which is consumed [26]:

$$\lim_{t\to\infty} \int_0^t \left(P_{\text{harvest}}(t) - P_{\text{node}}(t)\right) \text{dt} \geq 0 \tag{3}$$

However this condition doesn't account for the fact that V_{store} could fall below V_{kill}, so a second condition ensures that over any interval of time the system won't consume so much energy that it causes itself to fail from $V_{\text{store}} < V_{\text{kill}}$:

$$\forall t, \int_0^t \left(P_{\text{node}}(t) - P_{\text{harvest}}(t)\right) \leq \frac{1}{2} C_{\text{store}} \left(V_0^2 - V_{\text{kill}}^2\right) \tag{4}$$

Where V_0 is the initial voltage on C_{store} at $t = 0$. Intuitively, (4) reflects the possibility for the node to consume high amounts of power that exceed P_{harvest} for a brief amount, depleting any usable amount of stored energy, as long as V_{store} doesn't fall below V_{kill}. Practically, this condition is leveraged to transmit (TX) and receive (RX) data, since the instantaneous power consumed during active wireless operation is much higher than any other part of an IoT node.

Fig. 1.3(a) shows an example transient operating diagram of an IoT node, where the node is in a minimally-functional standby state (P_{node}=P_{standby}) for most of the time that could include a timer circuit, a wakeup radio, and some background digital computation. The

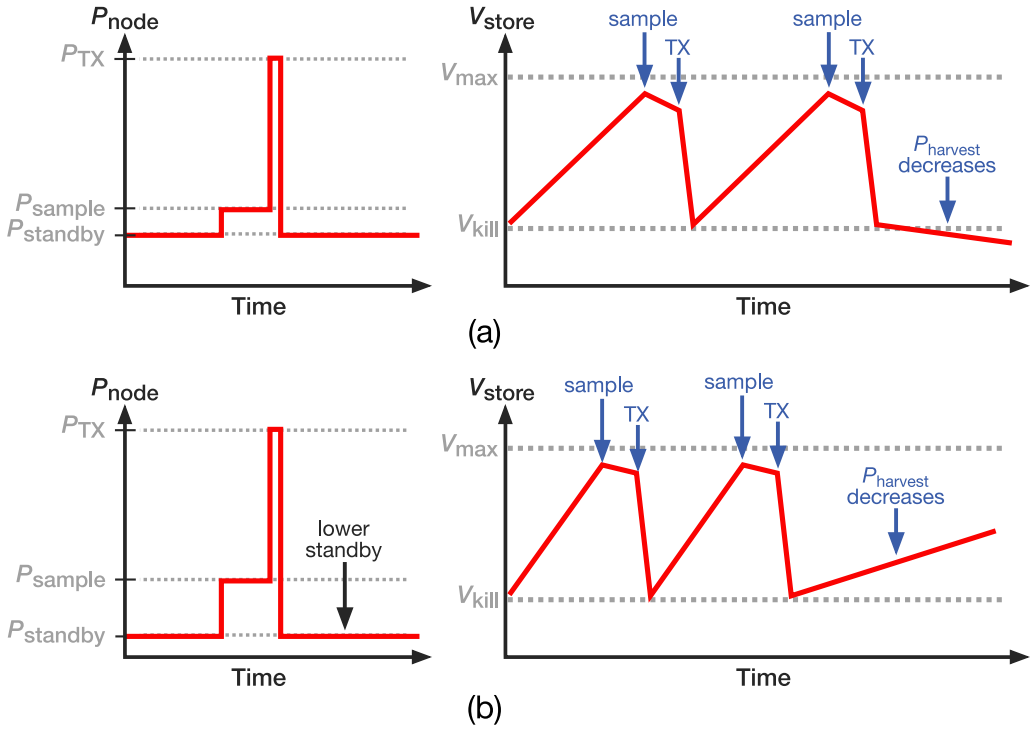

Fig. 1.3. Transient operating conditions for an IoT node that waits in a standby mode until enough energy has been harvested and stored that it can enable high-power circuits to obtain, process, and transmit a data sample.

node will periodically wakes up to obtain and process a data sample, consuming a relatively higher amount of power (P_{node}=P_{sample}), and then finally enter the highest-power mode to transmit that data (P_{node}=P_{TX}). If we assume that $P_{harvest}$ is constant and lies somewhere between $P_{standby}$ and P_{sample}, then V_{store} will be replenished to V_{max} while the node is in standby (meeting (3)), and then discharge rapidly during the sample and transmit phases until the usable stored energy is depleted (limit of (4)). The power required to wirelessly transmit data through free space is a physically-unavoidable cost that will almost always require a significantly high amount of instantaneous power that leverages (4) (10's of pJ/bit are typical), but it has been shown that the associated circuits can also be designed to reach extremely low standby power levels (less than 100pW) to ensure that they make a negligible contribution to $P_{standby}$ [27]. For the remainder of this discussion, it will be helpful to view the TX power as a fixed energy cost. Now, the focus for this dissertation turns to $P_{standby}$, which plays a critical role in the overall operating ability. However much $P_{standby}$ can be reduced below $P_{harvest}$, the faster C_{store} will be recharged and ready to transmit again. An example waveform of the same system operating with lower $P_{standby}$ is shown in Fig. 1.3(b). A key

observation here reflects (3) by showing that if $P_{harvest}$ decreases beneath $P_{standby}$, the node will eventually die and not be able to operate at all. A system with lower $P_{standby}$ (Fig. 1.3(b)) can therefore sustain always-on self-powered operation from smaller amounts of harvested power. As shown by Table 1, several sources exist that could enable always-on self-powered operation if $P_{standby}$ can be pushed down into the pW-nW range.

2 Digital Logic

2.1 Introduction

At the highest level of abstraction, digital circuits can be viewed as black boxes that encapsulate some algorithm or function based on a set of binary inputs. On a more detailed level of abstraction, digital circuits concern the design of electronic circuits known as logic gates that implement Boolean algebra functions which are essential for computing on the aforementioned binary inputs. There are many ways to arrange transistors to achieve this goal that can generally be grouped into logic "families" or "styles" based on unique aspects of their topology or implementation. Notable examples include domino logic and pass-transistor logic, and [28, 29] provide good starting points to explore the vast number of alternatives. An interesting discussion on the general electrical requirements for implementing these types of logic gates is given by A.W. Lo in [30]. A single logic family, traditional static-CMOS logic [31], has reigned supreme for many decades since its inception due to its balanced performance in critical performance metrics such as area, robustness, speed, and power.

Regardless of the logic family under consideration, power consumption in digital circuits (and really any other type of circuit) can be represented by the sum of the dynamic and static power:

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}} \tag{5}$$

Historically, the dynamic power consumption P_{dynamic} is most significant in digital circuits, which is caused from the charging and discharging of gate and wire capacitances throughout the circuit on each clock cycle. It is quadratically sensitive to the supply voltage V_{dd} and linearly dependent on the clock frequency f_{clk}, gate capacitance per transistor C_{gate}, total number of transistors N, and activity factor α of each transistor:

$$P_{\text{dynamic}} = \alpha f_{\text{clk}} N C_{\text{gate}} V_{\text{dd}}^2 \tag{6}$$

Note that wire capacitance can either be neglected for simplicity in this model or lumped in with C_{gate} if careful calibration is used based on physical parameters of the implemented circuit. Differences in logic families are generally reflected in this model by different N values, which would indicate more or less transistors being required to form each logic gate.

Static power P_{static} is always present due to various leakage currents in each transistor such as subthreshold leakage, gate tunneling leakage, and junction leakage. These will be discussed next in section 2.1.1. These currents can be easily represented as a total quantity I_{leak} which represents the total leakage current per transistor and is ultimately still a function of device parameters such as the transistor size (width and length, which affect C_{gate}) and the supply voltage V_{dd}. An easy-to-use expression for P_{static} is then:

$$P_{static} = I_{leak} N V_{dd} \tag{7}$$

It follows from (6) and (7) that dynamic power can be greatly reduced by using lower

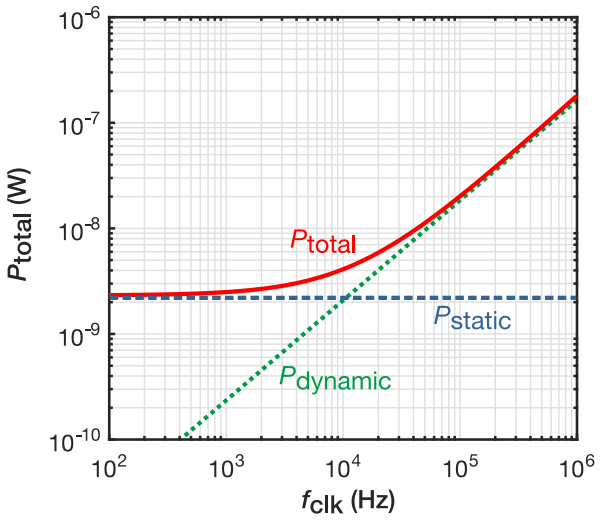

Fig. 2.1. Power consumption versus operating frequency (f_{clk}) of a 32-bit ripple-carry adder in 65nm, built with traditional static-CMOS logic gates, while operating at a 0.5V supply voltage V_{dd}.

supply voltages and then decreasing the clock frequency as needed, which could completely eliminate the dynamic power consumption once f_{clk}=0Hz. Note that the use of both of these techniques simultaneously is referred to as Dynamic Voltage and Frequency Scaling (DVFS) [32, 33] and is one of the more fundamental techniques for runtime *dynamic* power management. Depending on the overall architecture of the digital circuit (N, C_{gate}, α) and

operating parameters (V_{dd}, f_{clk}), $P_{dynamic}$ can fall below P_{static}, as shown in Fig. 2.1, meaning that the total power will be limited by various leakage currents that comprise the static power. This issue has only recently begun to bottleneck digital circuit power reduction efforts due to the proliferation of very-low (i.e., subthreshold) supply voltage operation and the advent of IoT applications that can tolerate very low (Hz to kHz-range) operating frequencies.

2.1.1 Leakage Current

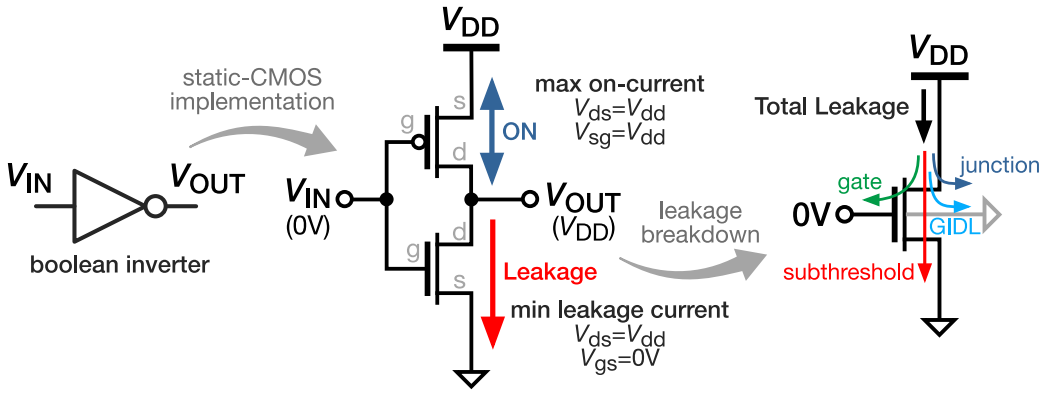

Fig. 2.2. Traditional static-CMOS implementation of a Boolean inverter with illustration of leakage currents that travel through the NMOS device.

In general, some amount of leakage current will flow between any two terminals (drain, source, body, or gate) of a transistor where a voltage differential is present. As far as power consumption is concerned, it is easiest to consider total leakage current per transistor as the sum of leakage currents traveling through the its drain, as shown in Fig. 2.2 and Fig. 2.3. Leakage through the gate can be caused by many underlying mechanisms, but in Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET)s the two most important mechanisms are Fowler-Nordheim tunneling and direct tunneling [34–36]. Fowler-Nordheim tunneling occurs when electrons in the conduction band of the substrate pass into the conduction band of the gate oxide when high gate voltages (usually several Volts) are present. Direct tunneling is more quantum-mechanical in nature and is usually much more significant than Fowler-Nordheim tunneling in devices with oxides that are less than 4nm thick (note that equivalent oxide thickness, or dielectric constant, does not matter here). Direct tunneling occurs when charge carriers tunnel directly through the gate oxide and into the energy bands

of the gate contact due to the extremely small physical separation between the two voltages. The current can tunnel from the gate directly into the source or drain, or into the channel where it can be swept into the source or drain or leave through the substrate. An expression for gate-to-drain or gate-to-source tunneling current can be given by [37]:

$$I_{gd} = W_{eff} k_1 V_{gd}^2 e^{(-m(A-B\times V_{gd})(1+C\times V_{gd}))} \tag{8}$$

$$\approx W_{eff} k_2 V_{gd}^2 \tag{9}$$

Where W_{eff} is the effective width of the transistor and k_1 is a scaling factor to account for technology. The exponential term contains many constant values to provide more accurate results at high gate voltages (V_{gd}), but this term can be reduced to a constant value and combined with k_1 to produce a new scaling factor k_2 that yields a simpler model that is equally as accurate at voltages less than 1V or so. Note that this applies to all voltage polarities, with current being able to flow into or out of the gate depending on the gate and source/drain voltages. Direct tunneling current I_{gc} into the channel (which then flows through the source, drain, or body) is typically greater than direct gate-drain and gate-source tunneling current due to the increased physical area over which these processes occur. Whereas (9) only depends on the transistor width, the gate-channel current depends on gate area:

$$I_{gc0} = W_{eff} L_{eff} k_1 V_{gs} V_{aux} e^{(-m(A-B\times V_{gs})(1+C\times V_{gs}))} \tag{10}$$

$$\approx W_{eff} L_{eff} k_2 V_{gs} V_{aux} \tag{11}$$

where both (10) and (11) are defined at $V_{ds} = 0\,V$ and have slightly modified behavior if if V_{ds} is nonzero [37]. Note that the constant values A, B, C, k_1, and k_2 do not correlate with those used in (8) and (9). A key difference in I_{gc} from I_{gd} and I_{gs} is temperature, which is reflected in V_{aux} by

$$V_{aux} = \frac{kT}{q} \log\left(1 + e^{q(V_{gs}-V_{th0})/kT}\right) \tag{12}$$

When high drain voltages are present and a low (especially negative) gate voltage exists, a phenomenon called GIDL occurs due to vertical overlap between the gate and drain that creates a buried depletion pocket at the edge of the drain near the surface of the silicon that

allows electrons to travel between the drain and the substrate via band-to-band tunneling. There is not a very convenient expression for GIDL current, but thankfully it isn't needed for this dissertation since it is negligible at all but the most extreme operating points of the MOSFET.

Junction leakage occurs when current travels through the reverse-biased PN junctions in various areas of the transistor. The most common areas for this to occur is in the junctions formed between the source and the substrate and the drain and the substrate. The expression for this is simply based on diode current:

$$I_{bs} = I_{S,bs}\left(e^{\frac{qV_{bs}}{kT}} - 1\right) \tag{13}$$

where I_S is the reverse-bias saturation current, k is Boltzmann's constant, and T is the temperature.

Subthreshold leakage is traditionally the most significant, especially at the voltages and operating points that transistors typically see when they are implemented in traditional static CMOS logic gates. It occurs due to a fundamental inability of the transistor channel to fully cut off the thermally-drive diffusion of electrons between its source and drain as long as a voltage exists between across its source and drain terminals. An expression for the subthreshold current is given by

$$I_{ds} = I_0 e^{\frac{(V_{gs} - V_{th0} + \eta V_{ds} - k\gamma V_{bs})}{n_0 v_T}}\left(1 - e^{\frac{-V_{ds}}{v_T}}\right) \tag{14}$$

Where V_{th0} is the nominal threshold voltage, η_0 is the Drain-Induced Barrier Lowering (DIBL) coefficient, η_B is the body effect coefficient, n_0 is the nominal subthreshold swing ideality factor that defines the "slope" of the subthreshold current characteristic, v_T is the thermal voltage which is equal to kT/q, and I_0 is the current at threshold ($V_{gs} = V_{th0}$) that may be approximated by

$$I_0 = \mu_0 C_{ox} \frac{W}{L}(n-1){v_T}^2 \tag{15}$$

Where μ_0 is the charge carrier mobility and C_{ox} is the gate oxide unit capacitance. The final term of (14) models the fact that no subthreshold current will flow if the drain-source voltage is zero.

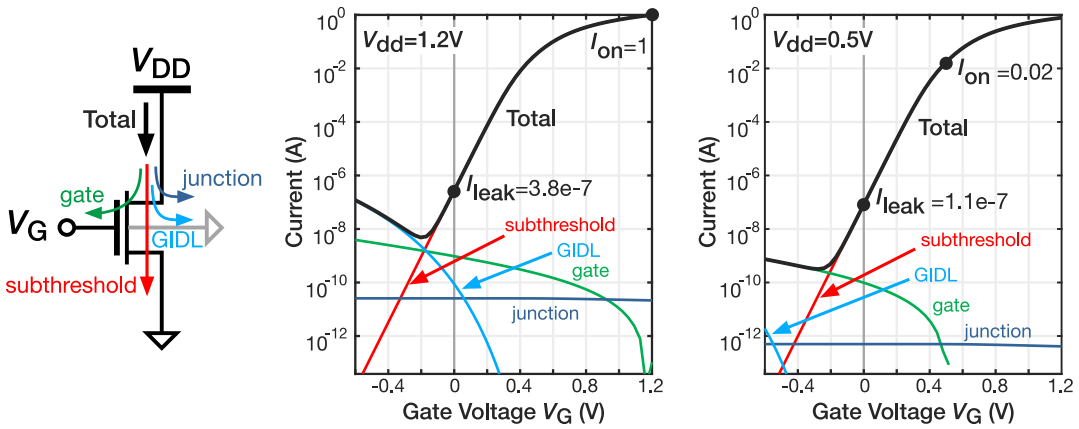

Fig. 2.3. Breakdown of drain current in an NMOS transistor as it would be used in a digital logic gate. Total drain current versus gate voltage is shown for supply voltages of 1.2V and 0.5V. At negative gate voltages, either GIDL or gate leakage will cause an increase in total current, depending on the supply voltage.

The total drain current of an NMOS transistor as a function of the gate voltage V_g is shown by Fig. 2.3. When a V_{dd} of 1.2V is assumed, a gate voltage of 1.2V puts the transistor into the saturation region where a high amount of on-current (I_{on}) is delivered. As the V_{gs} decreases, the transistor enters the sub-threshold region where the total current begins to taper off exponentially. Note the overwhelming dominance of the subthreshold current in this region, which continues to decrease at the same rate V_{gs} drops below 0V. As V_{gs} becomes more negative, however, gate leakage rises due to the growing voltage between the 1.2V drain and the negative gate. Moreover, GIDL is significant here again due to the high drain voltage. It is important to observe that in traditional static CMOS logic gates, the transistors never see a gate voltage below 0V (Fig. 2.2). Therefore, the total leakage current I_{leak} is defined by the total drain current at V_{gs}=0V, which is again overwhelmingly dominated by subthreshold leakage. At this point, a simplified model can be given for subthreshold leakage current, by applying this knowledge to (14), along with the assumption that the logic gates use standard body-biasing approach (no forward or Reverse Body Bias (RBB)) and a supply voltage that is above a few multiples of the thermal voltage:

$$I_{sub-leak} = I_0 e^{\frac{-V_{th0}+\eta V_{dd}}{n_0 v_T}} \tag{16}$$

Reducing V_{dd} from 1.2V to 0.5V will reduce many of the leakage currents. GIDL becomes essentially negligible, and gate leakage also decreases due to the lower drain voltage.

Subthreshold current decreases due to DIBL, which is represented by the linear scaling term η in 14. This yield a modest (3x) decrease in total leakage current, but also limits the maximum gate voltage in a traditional static-CMOS logic gate to 0.5V ($=V_{dd}$), which sharply reduces the on-current (I_{on}) by about 50x.

2.1.2 Leakage Reduction for Digital Circuits

Efforts to reduce power consumption in digital circuits focused originally on reducing supply voltage V_{dd} and clock frequency f_{clk} to reduce dynamic power [29]. This scaling reached the subthreshold regime around 2 decades ago [38], where circuits are intentionally operated with $V_{dd} < V_{th}$ and low clock frequencies in the kHz-range to keep dynamic power low. In cases where subthreshold digital circuits need further dynamic power reduction, clock gating is used to effectively duty-cycle the dynamic power consumption by temporarily disabling the active operation of the circuit [39, 40].

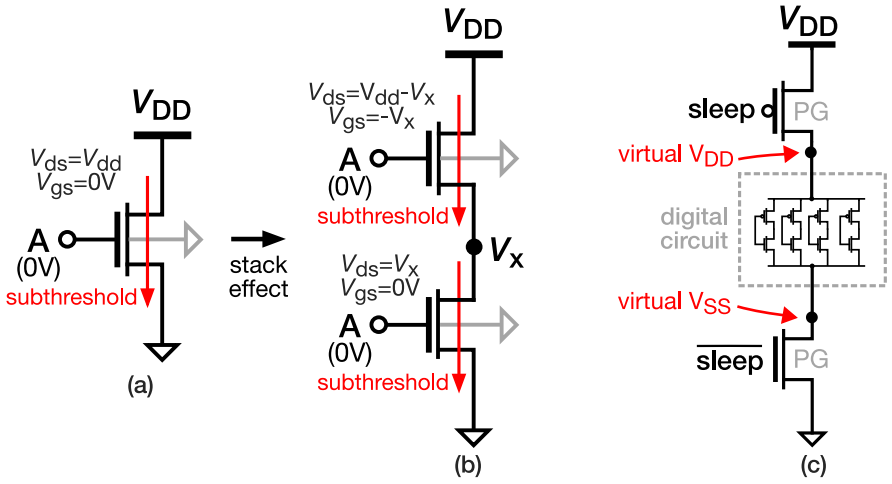

Fig. 2.4. Illustration of the stack effect for reducing subthreshold leakage and its application in power gating to cut off subthreshold leakage current to a digital circuit.

Presently, with dynamic power firmly under control, further power reduction efforts must target the static power consumption caused by leakage currents. As discussed previously in 2.1.1, subthreshold leakage is the most significant in traditional static CMOS logic gates. Prior techniques for reducing subthreshold leakage have focused on reducing the drain-source potential (V_{ds}) of each leaking transistor which both linearly reduces the power consumption due to Ohm's law but also reduces the subthreshold current itself due to DIBL. The most

obvious way to achieve this is by using lower supply voltages. The theoretical minimum supply voltage for static-CMOS logic has been calculated as 36mV [41], which is based on von Neumann's thermodynamically-driven estimate on the minimum amount of energy required to perform an elementary and reversible act of computation, which was later proven by Landauer by considering a "computer" that uses a bistable quantum well to shift the position of a single electron between adjacent wells that represent the two possible states for a piece of binary data [42, 43]. In practice, non-idealities of static-CMOS circuits such as noise and leakage limit the minimum supply voltage to a much higher value, closer to 100mV. To deal with these practical limitations, a Schmitt trigger-based logic style was proposed to enable ultra-low voltage operation at nearly 60mV [44].

In addition to supply voltage reduction, which often is already performed for dynamic power reduction, there are some more transistor-oriented techniques for further reducing the V_{ds} of leaking transistors. For example, the single "off" transistor that has been demonstrated so far, which is shown again in Fig. 2.4(a), sees a $V_{ds} = V_{dd}$ in its natural leaking state. If this transistor is broken into two series transistors (this is still functionally identical), an intermediate voltage node V_x is formed that will float to an intermediate value $0 < V_x < V_{dd}$ due to contention in the leakage currents between the two "stacked" transistors that acts as a voltage divider [45, 46]. This is shown in Fig. 2.4(b). Both transistors will see a V_{ds} less than V_{dd} now, and the upper transistor will even see a negative V_{gs} which further reduces its subthreshold current (see (14)). Many static-CMOS logic gates naturally have stacked transistors, such as a NAND2, so this effect can be leveraged by forcing the inputs to each gate to the necessary voltages (both A and B equal to 0, in this example) when a non-functioning sleep state is desired.

Power gating is a more dedicated method for enabling a low-leakage sleep state, which is based on the stack effect [47]. It works by inserting power-gating (PG) transistors in series between the digital circuit and the original voltage rails that act as switches to cut off the internal digital circuit from the supply rails as shown in Fig. 2.4(c). When the PG transistors are disabled by providing a logic-high `sleep` signal, the circuit becomes non-functioning and the internal "virtual" supply voltage rails float to intermediate voltages. Sizes of the PG transistors should be carefully picked to ensure that adequate current is supplied to the

circuit during active operation [48], and in some cases the PG transistors can be driven to negative gate voltages (overdriven by positive voltages, the case of a PMOS transistor) to help suppress their subthreshold current in the sleep state [49].

Other transistor-level subthreshold leakage reduction techniques focus on making the channel less accommodating to the diffusion current in the off-state. This can be done by applying a reverse bias voltage to the transistor body to increase the depletion regions which thereby increases the effective threshold voltage required to invert the channel [50]. Lengthening the channel of the transistor can reduce the occurrence of dopant interactions that lead to an effective short-circuit between the source and drain, and also adds physical distance between the source and drain that increases the energy required for an electron to traverse the channel, which is again effectively raising the threshold voltage [DT22, 51, 52].

2.1.3 DLS Logic

A more recent effort to reduce static power is a new logic style that improves on traditional static-CMOS by intrinsically creating a negative gate voltage on each leaking transistor which results in significantly reduced subthreshold leakage current as shown in Fig. 2.3. The transistor implementation for this approach is shown in Fig. 2.5 with comparison to traditional static-CMOS logic. This technique interestingly acts in a similar way to the power-gating technique shown in Fig. 2.4(c). Rather than using manually-controlled PG transistors to disable an entire digital circuit, this technique is essentially inserting the PG transistors into each gate (M_{nh} and M_{pf}) and allowing each gate to individually control its own power-gating sleep mode based on its own output voltage which is deterministic based on our knowledge that the gate is inherently inverting in its functionality. In contrast with the block-level PG technique, this method only cuts off a subsection of each logic gate (either the pull-up or pull-down network) so that the remainder of the gate can remain fully operational. This mechanism made an appearance first as part of a diode-based SRAM [53], and was later formally invented as a standalone “Ultra-Low Power” (ULP) logic style supported by simulation data [54]. While silicon-proven experimental data later became available from ring oscillator test structures in 130μm [55], the style did not gain significant attention until several years later when it was used to experimentally demonstrate a full microprocessor in 180μm and

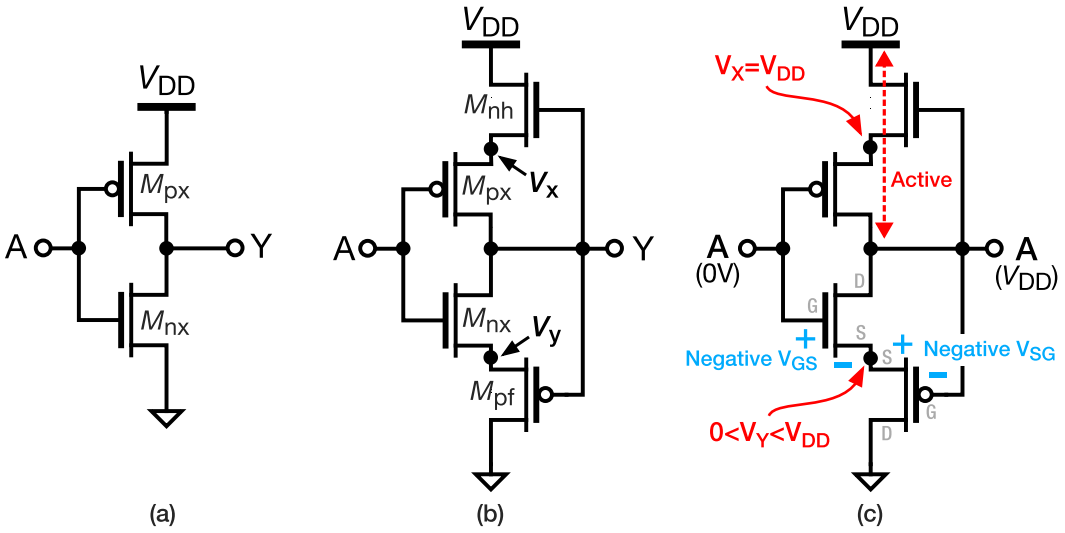

Fig. 2.5. (a) Traditional static-CMOS inverter, (b) Dynamic Leakage Suppression (DLS) inverter, (c) Illustration of leakge reduction behavior in a DLS inverter.

renamed as “Dynamic Leakage Suppression” (DLS) logic [1] without a link to the original design.

These works provide a few interesting insights on the operation and performance of DLS logic, but many of its aspects are still poorly understood. First, it is interesting to note that the existing studies on DLS logic take place within technology nodes that see relatively less leakage currents than modern deep sub-micron nodes, especially with respect to gate leakage due to the use of thicker oxides. A simplified analytical modeling of the total leakage current in DLS logic is therefore able to consider only subthreshold leakage while assuming that other forms of leakage are negligible [56]. The results suggest that total leakage current of DLS logic will perpetually decrease as supply voltage increases due to this creating a more negative gate voltage on the leaking transistors that reduces subthreshold leakage strongly enough to out-scale DIBL and Ohm’s law. However, in certain conditions, total leakage can instead increase with supply voltage due to gate leakage exceeding subthreshold leakage. This result was reflected experimentally by [1] but not further analyzed. As a result, the application of DLS logic has been emphasized for situations where subthreshold leakage is naturally very significant, such as in high-temperature applications. As per the current state-of-the-art, a comprehensive analytical modeling and analysis of leakage power in DLS logic does not exist.

The active performance of DLS logic is also only partially understood. It is known that the active drive current of a DLS gate is limited by the header (M_{nh}) and footer (M_{pf}) transistors that never see a V_{gs} above 0V at any point of the gate's operation. This would suggest that the gates are limited to very slow operating frequencies, which was again experimentally confirmed by [1] which operated at a maximum of 15Hz, but was not mathematically assessed any further. General design knobs such as device selection and device sizing have only been briefly investigated as part of a DLS design methodology that focuses on header and footer sizing to ensure equal rise and fall times in the face of process variation. Again, due to the lack of analytical modeling, the ramifications of these design knobs on other aspects of performance such as leakage and delay is not currently understood. Moreover, there is not a good understanding on what the maximum performance limits of DLS logic could be based on these design knobs and on the underlying technology factors.

Most recently, a variation on DLS logic was introduced that creates a hybrid between traditional DLS and traditional static-CMOS [2]. This is achieved by adding an extra header and footer transistor in parallel to the existing M_{nh} and M_{pf} which can be controlled with an externally-supplied signal that allows M_{nh} and M_{pf} to be shorted on-command to dynamically reconfigure each gate as either DLS or static CMOS logic. However, an issue with this approach is a lack of continuous scalability between DLS performance and CMOS performance.

2.2 Modeling for DLS Logic

This section derives theoretical models for the key operating characteristics for DLS logic and provides experimental verification for these models to understand the technology and device-specific factors that influence DLS performance. First, the DC switching characteristics are analyzed to define operating margins that will ensure robust operation. Next, an analysis of power consumption in DLS-based circuits is performed and a theoretical expression that allows for power minimization is obtained. Operating speed is briefly addressed by identifying factors that affect gate delay. Finally, a review of the derived models enables the identification of major sources of variation in DLS logic. Together, the derived models and insights from this section help provide an understanding of the availability and strength of design knobs for tuning performance and ensuring robustness.

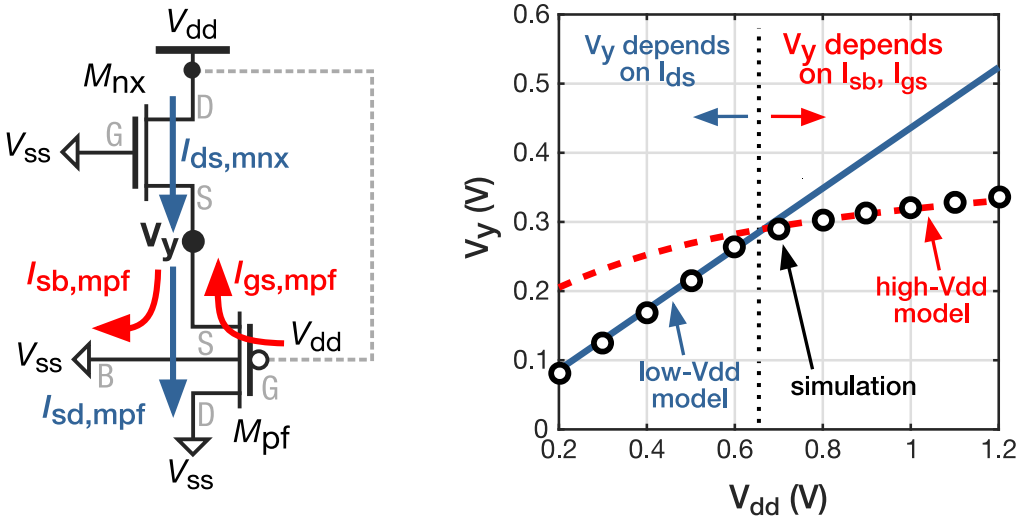

Fig. 2.6. Pull-down network of a DLS inverter with leakage current sources that affect the DC value of V_y. At low V_{dd}, contention between subthreshold currents determines V_y, while at high V_{dd}, it is contention between gate and junction leakage that determines V_y. Models for both cases agree well with simulated data.

2.2.1 DC Switching Characteristics

An important step before beginning the DC analysis is to first derive an expression for the internal V_x and V_y voltages that are pictured in Fig. 2.5(c). These voltages will end up being a key factor in the DC characteristics of the DLS gate. Depending on the input and output states of the DLS gate, either V_x or V_y will float to an intermediate voltage based on a variety

of stray leakage currents that travel in and out of the node. Again considering the case in Fig. 2.5(c) for a DLS inverter with an input A=0V, the equivalent pull-down network is now shown in Fig. 2.6 where the drain of M_{nx} is approximately equal to V_{dd} which follows as feedback to the gate of M_{pf}. The intermediate node voltage V_y would be solved by applying Kirchoff's Current Law (KCL) at V_y while considering the subthreshold, gate tunneling, and junction leakage contributions from each transistor. Based on the equivalent model shown, components can be identified in each transistor, which are displayed in Table 2. Solving KCL for all of these components would result in a complex solution, so the approach can be greatly simplified by first assessing what the major contributors are and limiting the application of KCL only to those major contributors.

Source	Expression	Key Voltage Source
Transistor Mnx		
Subthreshold current	$I_{ds,mnx}$	$V_{gs} = -V_y$ $V_{ds} = V_{dd} - V_y$
Gate leakage	$I_{sg,mnx}$	$V_{sg} = V_y$
Junction leakage	$I_{sb,mnx}$	$V_{sb} = V_y$
GISL	$I_{GISL,mnx}$	$V_{sb} = V_y$
Transistor Mpf		
Subthreshold current	$I_{sd,mpf}$	$V_{sg} = V_y - V_{dd}$ $V_{sd} = V_y$
Gate leakage	$I_{gs,mpf}$	$V_{sg} = V_{dd} - V_y$
Junction leakage	$I_{sb,mpf}$	$V_{sb} = V_y$
GISL	$I_{GISL,mpf}$	$V_{sb} = V_y$

Table 2. Leakage current sources that affect node V_y.

At relatively low voltages, junction leakage and gate leakage become negligible compared to subthreshold leakage, so V_y can be represented by contention in subthreshold leakage current between M_{nx} and M_{pf}. Using (14) and equating $I_{ds,mnx}$ and $I_{sd,mpf}$ yields an expression for V_y valid at low supply voltages:

$$V_{y,1} = \frac{n_{pf}\left(\eta_{nx} V_{dd} - V_{th,nx}\right) + n_{nx}\left(V_{dd} + V_{th,pf}\right)}{n_{pf}\left(1 + \eta_{nx} + k_{\gamma,nx}\right) + n_{nx}\left(1 + \eta_{pf} + k_{\gamma,pf}\right)} \tag{17}$$

where subscripts of 'nx' or 'pf' denote whether the parameter refers to M_{nx} or M_{pf}. It is useful to note here that if DIBL and the body effect are neglected M_{nx} and M_{pf} are assumed to be identical, then (17) becomes $V_{dd}/2$. Likewise, the steady-state voltage V_x in the PU network

can be found as

$$V_{x,1} = \frac{n_{px}\left(\eta_{nh} V_{dd} - V_{th,nh}\right) + n_{nh}\left(V_{dd} + k_{\gamma,px} V_{dd} + V_{th,px}\right)}{n_{px}\left(1 + \eta_{nh} + k_{\gamma,nh}\right) + n_{nh}\left(1 + \eta_{px} + k_{\gamma,px}\right)} \tag{18}$$

Depending on the technology parameters that dictate the relative significance of subthreshold leakage compared to the other forms of leakage current, (17) may be the only expression needed to represent V_y at any V_{dd}. However, in most other modern technologies, gate leakage will out-scale subthreshold leakage for DLS gates at higher supply voltages due to higher gate-source and gate-drain voltages being seen by each transistor. GIDL does not apply here since the induced drain-body current would not affect the V_y node. Instead, GISL would apply to this situation, but it is thankfully not a concern due to the V_y node being naturally limited to voltages that are usually no greater than $V_{dd}/2$, resulting in an insignificant amount of source-body current, similar to the DIBL result shown in Fig. 2.3 for $V_{dd} = 0.5\,V$. At this point, the simplified approach to solve for V_y at high V_{dd} would be to equate $I_{gs,mpf}$ with $I_{sg,mnx}$. One more issue remains, though: A previously unspecified aspect of DLS gate design is the potential difference in sizing and threshold voltge between different transistors. Most DLS implementations choose the header and footer transistors M_{nh} and M_{pf} to be larger and have lower threshold voltages than the internal transistors M_{px} and M_{nx} to improve output voltage levels, which will be discussed later in this chapter. In some cases, M_{pf} may also use a Forward Body Bias (FBB) to lower its threshold voltage even further, which leads to higher junction leakage. The result for this situation is that $I_{sg,mnx}$ can be several times lower than $I_{gs,mpf}$ due to the transistor size difference, and that the source junction of M_{pf} is large enough (and biased strongly enough) that $I_{sb,mpf}$ will pull a significant amount of current out of the V_y node. Therefore, for this work, the representation of V_y at high V_{dd} considers contention only between $I_{gs,mpf}$ and $I_{sb,mpf}$ as shown in Fig. 2.6. The resulting expression is found:

$$V_{y,2} = V_{dd} - 2v_T W_{-1}\left(-\frac{1}{2v_T}\sqrt{\frac{I_{s,pf} e^{V_{dd}/v_T}}{k_{0,pf}}}\right) \tag{19}$$

where W_{-1} is the negative or "lower" branch of the Lambert function, which is defined by $W(x) \times e^{W(x)} = x$. The resulting models of V_y at low V_{dd} and high V_{dd} are shown in Fig.

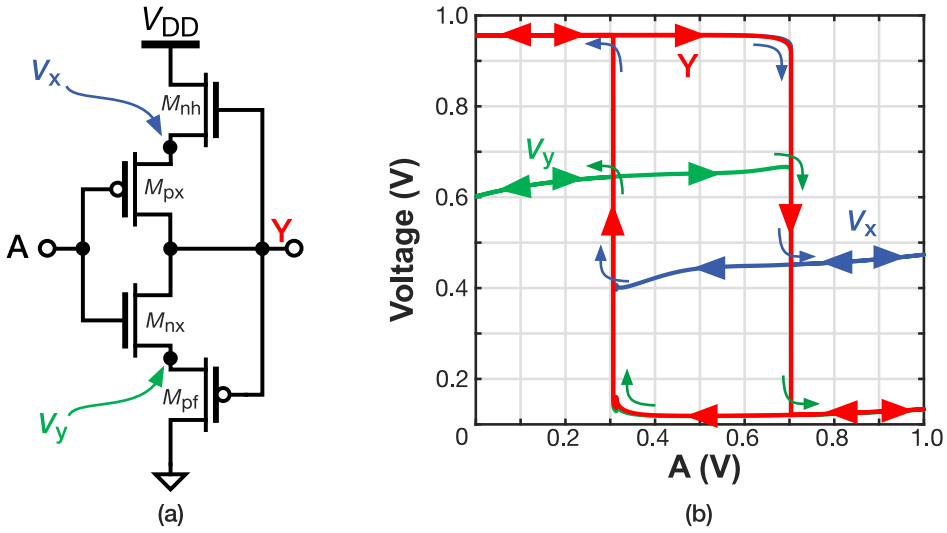

Fig. 2.7. Schematic of a DLS inverter and its simulated DC transfer characteristic, showing the output voltage Y as well as the internal nodes V_y and V_x.

2.6 and compared with simulation for a DLS inverter that uses minimum-size HVT inner transistors and 1μm-wide Low Threshold Voltage (LVT) transistors for M_{pf} and M_{nh}. Again, for V_x, the same approach can be taken to find

$$V_{x,2} = V_{dd} - 2v_T W_{-1}\left(-\frac{1}{2v_T}\sqrt{\frac{I_{s,px}}{k_{0,px}}}\right) \tag{20}$$

At this point, the DC characteristics of the DLS inverter, shown again in Fig. 2.7(a) for convenience, can be analyzed. Beginning in a state where the input A is 0V, the output is assumed to be logic-high at voltage $Y \approx V_{dd}$. At this DC operating point, the gate voltages of M_{nh} (=V_{dd}) and M_{px} (=0V) allow current to flow though the PU network. Once the Y and V_x nodes are charged, M_{nh} will settle in the cutoff region with $V_{gs} = 0\,V$ and M_{px} will settle in subthreshold with $V_{gs} > 0\,V$, although both transistors see a drain-source voltage of 0V. Meanwhile, the gates voltages of M_{nx} (=0V) and M_{pf} (=V_{dd}) prevent all but the smallest leakage currents from flowing, as was just shown in the derivation of V_y, placing them in a "super-cutoff" region where $V_{gs} < 0\,V$ ($V_{sg} < 0\,V$ for PMOS).

As A increases, it strengthens the drive current of M_{nx} by pushing it from super-cutoff toward standard cutoff, while also weakening the drive current of M_{px} by pulling it from subthreshold toward cutoff. Eventually, once A is high enough, M_{nx} overpowering M_{px} will

interrupt the feedback from the header and footer transistors and allow the gate to flip its output state. However, due to the large initial gap in operating regions between these two internal transistors (super-cutoff versus active subthreshold), it takes a reasonably large voltage on A (greater than $V_{dd}/2$) to allow M_{nx} to overpower M_{px}. Therefore, while the increasing A voltage does slightly increase V_y due to both increased subthreshold leakage and gate-source leakage through M_{nx}, it is not enough to produce a noticeable change in the output voltage of the gate, which is maintained by feedback through M_{nh} and M_{pf}.

Once A reaches a sufficiently high value V_{ih} for M_{nx} to overpower M_{px}, V_y and Y are allowed to equalize which means that M_{nx} can be ignored and M_{pf} will be in standard cutoff mode with relatively stronger leakage current that tries to pull Y towards 0V. At the same time, the PU network has a relatively weaker drive strength since M_{nh} is already in cutoff and M_{px} moving closer to cutoff from subthreshold (V_x is assumed to be floating near its previous value of V_{dd}). By definition, at this defined input level A=V_{ih}, the initial leakage current through M_{pf} in cutoff is stronger than the series leakage of M_{nh} and M_{px} in cutoff (see the stack effect in 2.1.2), so the output node Y will be pulled to ground. This feedback then applies to M_{pf} by further increasing its leakage to pull Y to 0V even faster, while M_{nh} begins to enter the super-cutoff region. The subsequent decrease in leakage current from M_{nh} makes it harder for V_x to stay charged to V_{dd} and the subthreshold leakage from M_{px} tries to pull V_x towards Y (=0V), so V_x will decrease to an equilibrium voltage somewhere between V_{dd} and 0V that will ultimately be determined by the contention in current between M_{nh} and M_{px}. The final intermediate voltage for V_x can be solved using the same approach used for (17) and (19), and will cause M_{px} to end up in the super-cutoff region as well. The reduced super-cutoff leakage current from M_{px} and M_{nh} ensure that the PD network is overpowering the PU network.

An idealized DLS VTC is shown in Fig. 2.8 which plots the V_{out} versus V_{in} for low-to-high ($V_{in} = 0 \rightarrow 1$) and high-to-low ($V_{in} = 1 \rightarrow 0$) transitions. The voltage levels of the gate output (V_{oh} and V_{ol}) at the logic-high and logic-low levels are found by mirroring the VTC across an imaginary axis equal to $V_{out} = -V_{in} + V_{dd}$ and finding the y-coordinates at which the original and mirrored intersect. These values represent what the high and low-output voltages would converge to in an infinite series of inverters driving each other if the input to the first inverter

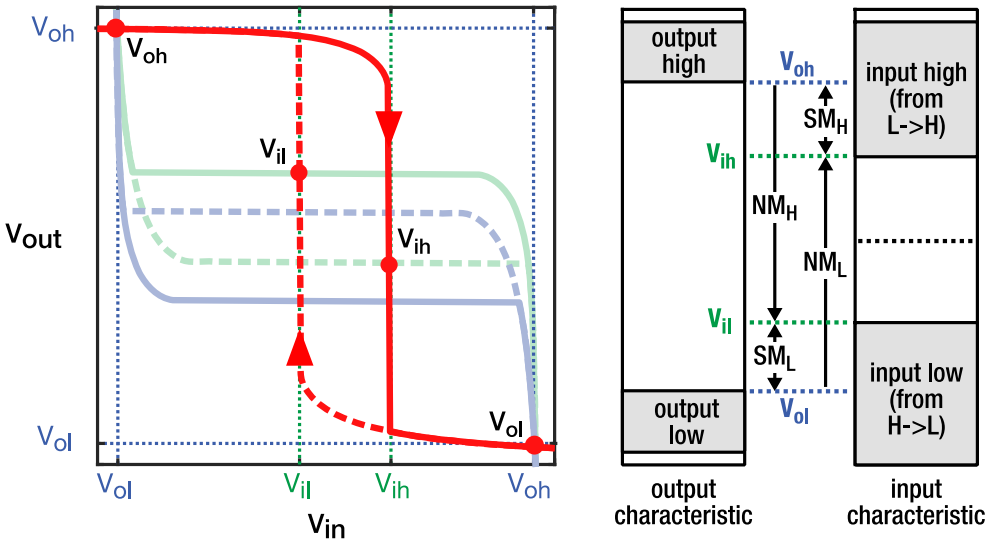

Fig. 2.8. Derivation of input threshold voltage levels and output voltage levels of the DLS gate, and illustration of how the noise margins (NM) and switching margins (SM) are derived from these values.

is ideal (V_{dd} or 0V), which mimics the behavior of a large digital circuit. In a standard inverter with symmetric $L \rightarrow H$ and $H \rightarrow L$ transitions, only one input switching threshold exists. Since the DLS inverter has asymmetric switching thresholds, a different input voltage is required in both the $V_{in} = 0 \rightarrow 1$ (V_{ih}) and $V_{in} = 1 \rightarrow 0$ (V_{il}) cases. To ensure proper operation of one DLS gate driving another, these voltages should be defined cyclically such that V_{ih} is the minimum rising V_{in} voltage required to make V_{out} drop below V_{il}, and vice versa for V_{il} being the maximum falling voltage required to make V_{out} increase above V_{ih}. This definition ensures that an infinite series of inverters would switch properly as long as input voltage to the first inverter crosses the V_{il} and V_{ih} values. These values are found by mirroring the VTC across an imaginary axis equal to $V_{out} = V_{in}$ and finding the x-coordinates at which the original and mirrored intersect.

These input-output characteristics give rise to several operating margins that help quantify the stability and robustness of the gate. The traditional example of this is the Noise Margin (NM) which defines how much noise can be tolerated at the input of a gate before its output is corrupted enough to affect the input of the next gate that it drives. Because the gate has two possible input states (logic high and logic low), two NMs exists for both cases depending on whether the input is high (NM_h) and low (NM_l). In an ideal gate with a single input switching threshold, if we assume $V_{oh} \approx V_{dd}$, $V_{ol} \approx$0V, and that the input threshold is symmetric

Expression	Name	Description	Derivation
Input-Output			
V_{ih}	High input threshold voltage	Minimum rising input voltage A required to make the output transition low (below V_{il})	Algorithm 1
V_{il}	Low input threshold voltage	Maximum falling input voltage A required to make the output transition high (above V_{ih})	Algorithm 1
V_{oh}	High output voltage level	Maximum output voltage from a DLS gate being driven by a logic-low (V_{ol}) input	Algorithm 2
V_{ol}	Low output voltage level	Minimum output voltage from a DLS gate being driven by a logic-high (V_{oh}) input	Algorithm 3
Operating Margins			
NM_h	High noise margin	The amount that a logic-high input (V_{oh}) to a DLS gate must decrease before it registers as a logic-low input (V_{il}) and causes the gate to invert its output	$V_{oh} - V_{il}$
NM_l	Low noise margin	The amount that a logic-low input (V_{oh}) to a DLS gate must increase before it registers as a logic-high input (V_{ih}) and causes the gate to invert its output	$V_{ih} - V_{ol}$
SM_h	High switching margin	The amount that a logic-high input (V_{oh}) to a DLS gate exceeds the high input threshold voltage necessary to invert the output of the gate	$V_{oh} - V_{ih}$
SM_l	Low switching margin	The amount that a logic-low input (V_{oh}) to a DLS gate falls beneath the low input threshold voltage necessary to invert the output of the gate	$V_{il} - V_{ol}$

Table 3. Definitions of the DC characteristics of the DLS logic gate.

($\approx V_{dd}/2$), then the theoretical maximum NM is also $V_{dd}/2$. Note that this theoretical limit applies to traditional static-CMOS logic and is quite high relative to most naturally-occurring noise sources (this is partly why static-CMOS logic is so popular), but its tendency to decrease along with V_{dd} is a key issue that limits the minimum V_{dd} for digital circuits. The separate V_{ih} and V_{il} of the DLS inverter allow both the NMs to exceed $V_{dd}/2$, providing the gate with higher than average immunity to noise. The high noise margin NM_h can be defined as $V_{oh} - V_{il}$, while the low noise margin NM_l can be defined as $V_{ih} - V_{ol}$. However, there is no "free lunch" here, as the tradeoff for a higher NM via a low $V_{il} < V_{dd}/2$ and a high $V_{ih} > V_{dd}/2$ is less voltage overhead left for the input signal to actually exceed these input threshold voltages. A new definition is now needed to define the Switching Margin (SM) as the voltage margin by which an input to a DLS gate exceeds the required V_{il} or V_{ih}. Therefore, the we can define $SM_h = V_{oh} - V_{ih}$ and $SM_l = V_{il} - V_{ol}$. The SM differs from the NM in that while NM defines how stable a gate is in retaining its current state, the SM defines how reliably (easily) a gate's state can be changed. Naturally, these quantities are complementary, where in an ideal gate with $V_{oh} \approx V_{dd}$ and $V_{ol} \approx 0$V, it can be said that $NM + SM = V_{dd}$. The DLS gate already has a

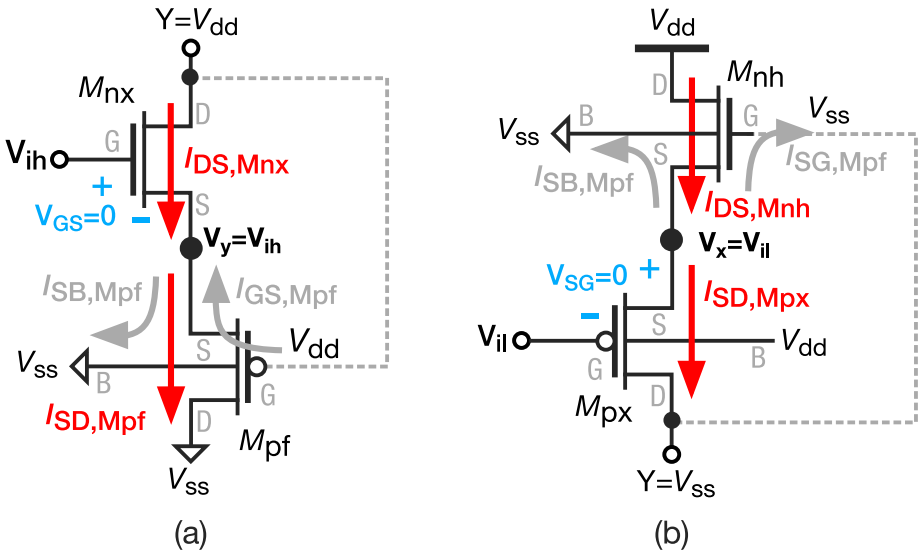

Fig. 2.9. Equivalent schematics for determining the input threshold voltages for a DLS logic gate. Part (a) shows the pull-down network to solve for V_{il}, part (b) shows the pull-up network to solve for V_{ih}.

high NM which puts the SM at risk to begin with, but once the non-ideal output voltage levels (V_{oh} and V_{ol}) are accounted for, the SM becomes a critical focus for the DLS gate to ensure proper operation. Table 3 summarizes the discussed input-output voltages and operating margins of the DLS DC characteristics.

Analytically, the simplest approach to solve for V_{ih} is to assume that M_{nx} will overpower M_{px} once its V_{gs} moves from being negative (super-cutoff) to positive (subthreshold). In other words, the (high) input threshold voltage for the DLS gate will be defined as the input required to set the gate-source voltage of M_{nx} to 0V, as shown in Fig. 2.9(a). This approach was originally identified in [54] and demonstrated as a point solution for a single V_{dd} in [57]. This calculation could be easily completed by assuming that V_y remains at its DC solution from A=0 (in (17),(19)) as A increases, since the solution for V_{ih} would then simply be equal to (17),(19). A slightly more accurate solution can be derived by using a similar approach as in (17),(19) while accounting for the fact that V_y is modulated by the rising input voltage A on the gate of M_{nx}. Due to the relatively high V_{gs} of M_{nx} at this operating point, it can assumed that its subthreshold current dominates its other sources of current (gate and junction). Assuming the subthreshold leakage is dominant in M_{pf} as well and that the output Y is still at V_{dd}, then the solution for V_{ih} can be solved from the contention between subthreshold currents flowing through M_{nx} and M_{pf}. Setting V_{gs} of M_{nx} to 0V and equating $I_{ds,mnx}$ with $I_{sd,mpf}$ and solving

for V_y yields the expression

$$V_{ih,1} = \frac{n_{pf}\left(\eta_{nx} V_{dd} - V_{th,nx}\right) + n_{nx}\left(V_{dd} + V_{th,pf}\right) + n_{nx} n_{pf} v_T \ln\left(\frac{I_{0,nx}}{I_{0,pf}}\right)}{n_{pf}\left(\eta_{nx} + k_{\gamma,nx}\right) + n_{nx}\left(1 + \eta_{pf} + k_{\gamma,pf}\right)} \tag{21}$$

It is interesting to note here that if DIBL and the body effect are neglected and we assume that both M_{nx} and M_{pf} have the same I_0 (V_{th0} can simply be re-defined to ensure that all devices have the save I_0) and subthreshold swing (n), we find that

$$V_{ih} \approx V_{dd} - V_{th,nx} + V_{th,pf} \tag{22}$$

Therefore, a loose design guideline is to ensure the ability to properly switch (V_{ih} must be less than V_{dd}) is that $V_{th,nx} > V_{th,pf}$. The same approach can also be used to find V_{il}, which is represented by Fig. 2.9(b), giving the following result:

$$V_{il,1} = \frac{n_{nh}\left(k_{\gamma,px} V_{dd} + V_{th,px}\right) + n_{px}\left(\eta_{nh} V_{dd} - V_{th,nh}\right) + n_{nh} n_{px} v_T \ln\left(\frac{I_{0,nh}}{I_{0,px}}\right)}{n_{nh}\left(\eta_{px} + k_{\gamma,px}\right) + n_{px}\left(1 + \eta_{nh} + k_{\gamma,nh}\right)} \tag{23}$$

Again, ignoring DIBL and the body effect and setting $I_{0,px} = I_{0,nh}$ and $n_{px} = n_{nh}$ shows us that

$$V_{il} \approx V_{th,px} - V_{th,nh} \tag{24}$$

Therefore, it should be also ensured that $V_{th,px} > V_{th,nh}$ so that the gate will be able to successfully operate. Still, in some cases when the subthreshold current of M_{pf} is low enough to be comparable to its gate and junction leakage currents, a different approach might be needed, just as in (19). For example, when solving for V_{ih}, the junction current of M_{pf} may be used instead of the subthreshold current, giving the following result:

$$V_{ih,2} = \frac{n_{pf}\left(\eta_{nx} V_{dd} - V_{th,nx}\right) + n_{nx}\left(V_{dd} + V_{th,pf}\right) + n_{nx} n_{pf} v_T \ln\left(\frac{I_{S,bs,pf}}{I_{0,nx}}\right)}{n_{pf}\left(\eta_{nx} + k_{\gamma,nx}\right) + n_{nx}\left(1 + \eta_{pf} + k_{\gamma,pf}\right)} \tag{25}$$

Fig. 2.10 shows the modeled and simulated values of V_{il} and V_{ih} versus V_{dd}, showing a good agreement. The low-V_{dd} model for V_{il} is valid across the fully supply voltage range,

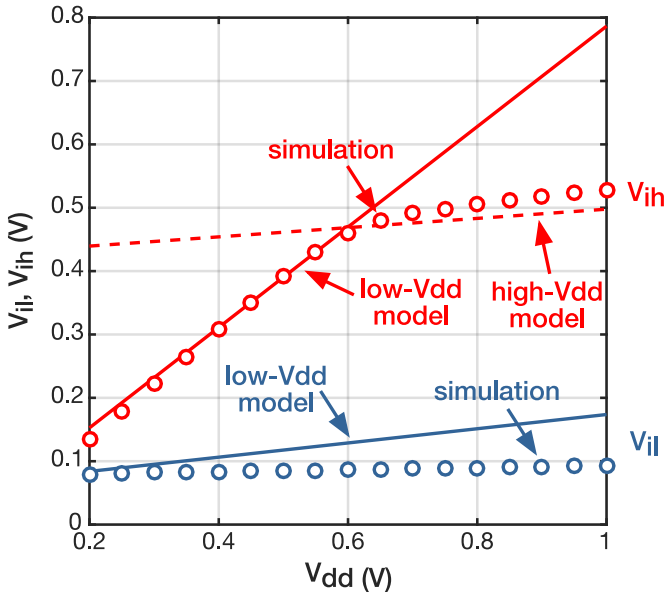

Fig. 2.10. Modeled and simulated input threshold voltages (V_{il} and V_{ih}) of the DLS inverter versus V_{dd}. Low-V_{dd} models are $V_{il,1}$ and $V_{ih,1}$ from (23) and (21), and the high-V_{dd} model for V_{ih} uses $V_{ih,2}$ from (25).

but V_{ih} requires the high-V_{dd} model from (25) after V_{dd} increases beyond 0.6V to maintain accuracy.

In static-CMOS logic gates, it is rare for the output voltage levels to deviate from the rail voltages (V_{dd} and V_{ss}). In an inverter, whichever network (pull-up or pull-down) that is turned on can typically source or sink enough current to keep the output level solidly fixed at the rail voltage despite any leakage currents that try to charge or discharge the output node. Of course, continuing the example for a static-CMOS inverter means the pull-up and pull-down networks are each just a single transistor. The ability for the "on" transistors in a logic gate to properly charge or discharge the output voltage of the gate is referred to as its "drive strength" – the more current these transistor can source or sink, the higher the drive strength. In a DLS gate, the drive strength provided by the "on" transistor is fixed by the subthreshold leakage of either M_{nh} or M_{pf}, as shown in Fig. 2.11. When the input A is equal to 0V, M_{px} is active such that is effectively shorted, so the drive strength from the pull-up network is equal to $M_{nh}$with its gate and source tied together (V_{gs}=0V). The subthreshold leakage of M_{nh} ($I_{ds,Mnh}$) should reasonably overpower the series leakage of M_{nx} and M_{pf} (as well as any other leakage sources) as mentioned during the analysis of the VTC such that V_{out} charges up to a relatively high value. However, if the relative difference between $I_{ds,Mnh}$ and the other

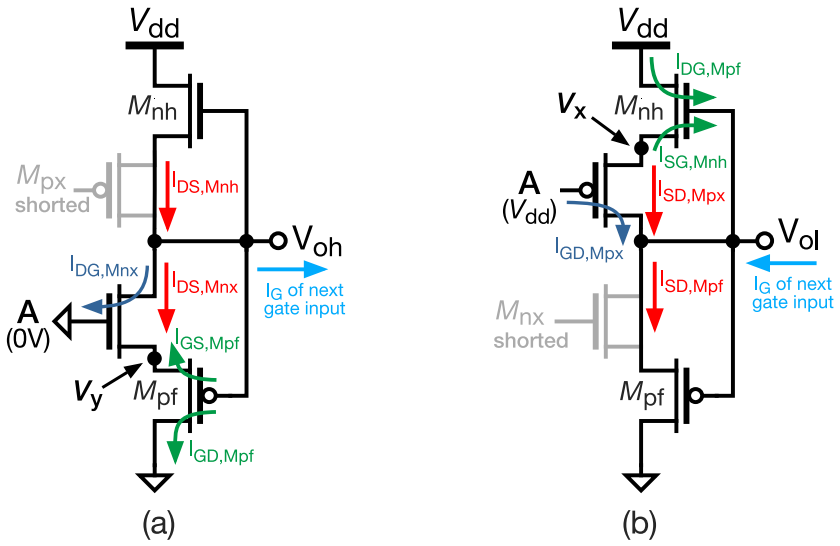

Fig. 2.11. Sources of drive current and parasitic leakages currents that determine the maximum output voltage V_{oh} (a) and minimum output voltage V_{ol} of the DLS gate.

sources of leakage is small enough, then the leakage sources will pull start slightly pulling the steady-state value of V_{out} down and M_{nh} simply won't have the drive strength to compensate for it. While an exact analytical expression would be helpful here, the very small V_{ds} of M_{nh} in the KCL expression requires a more complicated subthreshold current model to account for low V_{ds} (equation (14)) that leads to an intractable solution. Instead, a qualitative analysis here can just as easily provide insight into the factors that would degrade the output voltage swing of a DLS gate.

As shown by Fig. 2.11(a) for the case when A=0 and Y=V_{oh}, the leakage currents that try to discharge V_{out} come from the drain of M_{nx} (subthreshold and drain-gate leakage) and gate of M_{pf} (gate-source and gate-drain leakage). Additionally, there is a possibility for gate leakage to occur from the output node through the gates of the transistors in the input of the next logic gate. In other words, whereas it is common to consider the output load capacitance of a logic gate for delay and slew-rate purposes, the output parasitic leakage should also be considered for output voltage swing purposes. To assess V_{ol} when A=1 and Y=V_{ol} as shown in Fig. 2.11(b), the same situation is present where the leakage of M_{pf} tries to keep V_{out} properly pulled to 0V while stray currents come from the source of M_{px} (subthreshold and gate-source leakage) and gate of M_{nh} (drain-gate and source-gate) to charge V_{out}. Therefore, another loose guideline can be added here to ensure that the parasitic leakage currents do

not overpower the intrinsic drive strength from the active pull-up or pull-down transistors.

$$I_{ds,Mnh}\big|_{V_{gs}=0,V_{ds}=\sim kT/q} \gg I_{ds,Mnx}\big|_{V_{gs}=-V_y,V_{ds}=V_{dd}-V_y} + K_{gate}\,V_{dd}^2 \tag{26}$$

$$I_{sd,Mpf}\big|_{V_{sg}=0,V_{sd}=\sim kT/q} \gg I_{sd,Mpx}\big|_{V_{sg}=V_x-V_{dd},V_{sd}=V_x} + K_{gate}\,V_{dd}^2 \tag{27}$$

where $K_{gate}\,V_{dd}^2$ represents the sum of all gate leakage-based components by taking a pessimistic approach that assumes they all see a gate-source or gate-drain voltage of V_{dd} and K_{gate} is a general scaling term to represent the sum of all intrinsic gate leakages ($W_{eff}k_2$ of M_{nx}, M_{pf}, etc., as shown in (9)) as well as the fanout load. A V_{ds} of kT/q is chosen for the evaluation of M_{nh} and M_{pf} since technically their currents are equal to zero if V_{ds} is also set to its ideal value of zero.

2.2.2 Transient Characteristics

Analysis of the transient operating characteristics of the DLS gate builds on the previously-shown analysis for the DC VTC. A simulation of a falling-edge input (A=1→0) transient response is shown in Fig. 2.12 for a DLS inverter operating at V_{dd}=0.5V. To help illustrate the behavior of each gate during the transition, the time-axis is shown on a logarithmic scale. In the static state, before the input transitions low, Y is equal to 0V and V_x settles near $V_{dd}/2$ (exact value given by (18)), so M_{nh} is in super-cutoff with a negative V_{gs}. Likewise, since the gate of M_{px} is at V_{dd}, it also sees a negative V_{sg}. The input switches to a logic-low state, instantly putting M_{px} into the subthreshold regions and M_{nx} into the cutoff region. Since M_{px} is active, V_x and Y begin to equalize, which pulls them both slightly toward their average value (V_x droops and Y rises). This causes the V_{gs} of M_{nh} to converge to 0V, bringing it closer to the cutoff region, while M_{pf} is pushed into super-cutoff due to both Y rising toward $V_{dd}$and the internal node V_y settling at a value less than V_{dd}as in (17). By the time V_x and Y have equalized, Y is slightly above 0V, but still has to charge all the way to V_{dd} (technically V_{oh}) in the exact manner depicted in Fig. 2.11(a) where the subthreshold leakage of M_{nh} (minus the parasitic leakage) is responsible for supplying the drive current. The amount of time required for M_{nh} to charge Y is by far the most significant contributor to the overall transition delay, which could be modeled using a traditional t=CV/I delay that represents the

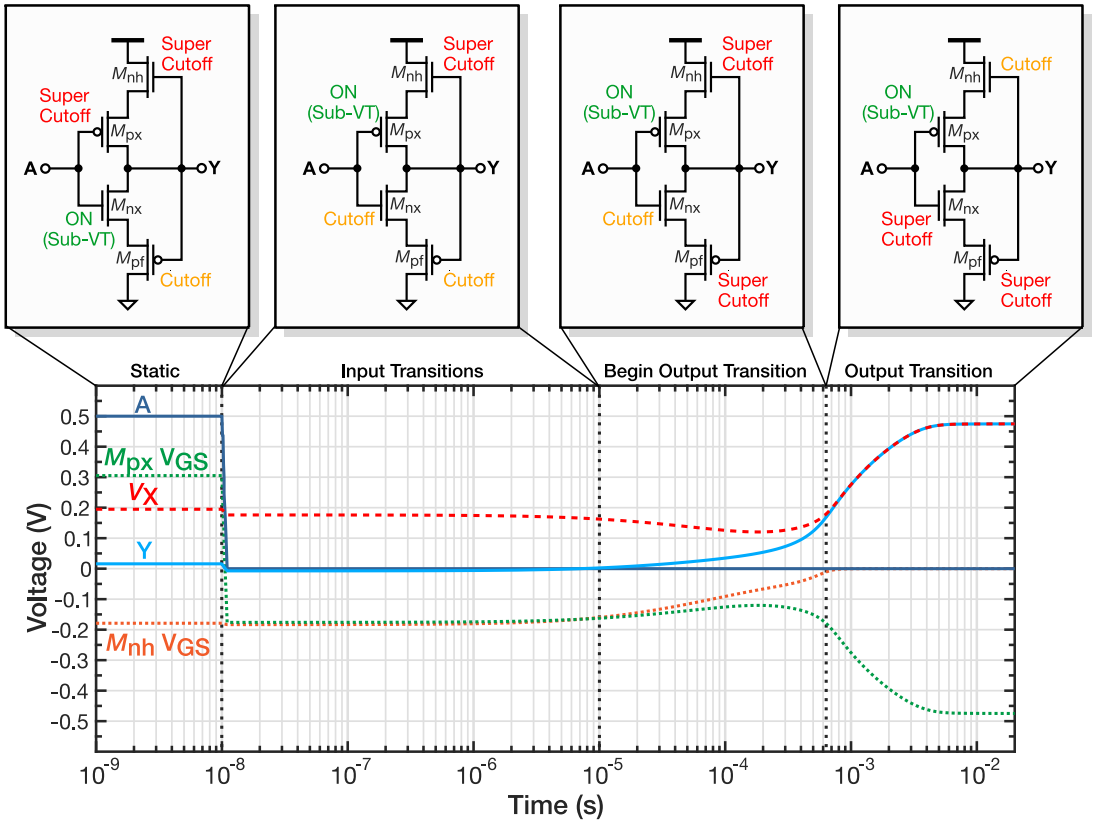

Fig. 2.12. Transient simulation of a DLS inverter switching from output low to output high at V_{dd}=0.5V. The operating regions of each transistor are illustrated at different key points throughout the transition.

amount of time required for a constant current I to charge a capacitor C to a voltage V. In this case, the capacitance (C_{gate}) is the self-capacitance of the DLS gate output (the gates of M_{nh} and M_{pf}) in addition to the input load capacitance of the next gate(s). If we assume that this capacitance begins charging at the point where V_x was in its steady-state (as in eq. (18)) and that it only needs to charge to V_{ih} in order to register at the input of the next gate, then we have

$$t_{gate} = \frac{C_{gate}\left(V_{ih} - V_{x,1}\right)}{I_{ds,Mnh}\big|_{V_{gs}=0, V_{ds}=V_{dd}-V_{x,1}} - K_{gate}V_{dd}^2} \tag{28}$$

Note that this model is based on the rising-edge output propagation delay. A unique falling-edge output propagation delay also exists, but it is assumed for simplicity at this point that both delays are equal. Because this model relies on the full representations of V_{ih} in (21) and V_x in (18), a full expression would become too complicated to provide any design insight. If we instead assume that the parasitic gate leakage is negligible (K_{gate}=0) and that C_{gate} must

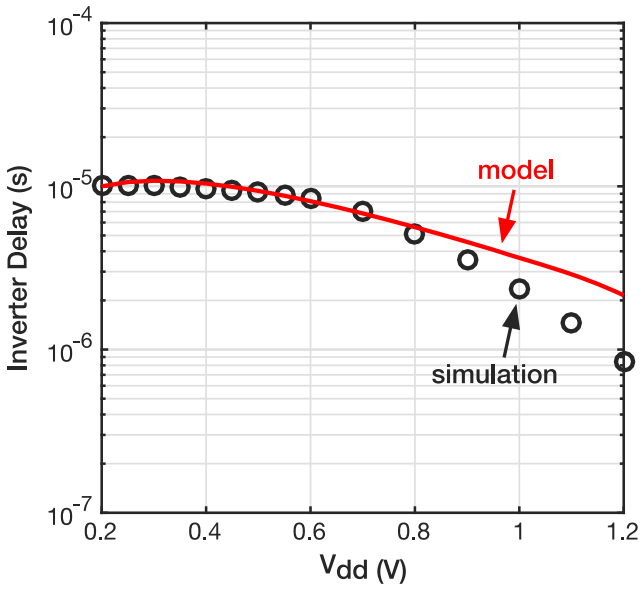

Fig. 2.13. Modeled and simulated gate delay (output low to high) of a DLS inverter versus supply voltage V_{dd}.

charge from 0V to V_{dd}, then $I_{ds,Mnh}$ can simply be evaluated at V_{gs}=0V and $V_{ds} = V_{dd}$, and the expression for gate delay becomes much more understandable:

$$t_{gate} \approx \frac{C_{gate} V_{dd}}{I_{0,nh} e^{\frac{\eta_{nh} V_{dd} - V_{th,nh}}{n_{nh} v_T}}} \tag{29}$$

The I in the CV/I delay, in the denominator of (29), now simply represents the leakage current of M_{nh}. Although the delay linearly increases with V_{dd} due to the numerator, the leakage of M_{nh} also increases with V_{dd} due to DIBL, which tends to outscale the linear numerator dependence. Therefore, the leakage of M_{nh} is the single biggest knob to control the gate delay, and the dependence of delay on V_{dd} depends on the push-pull between the increasing integration voltage (charging Y to a lower or higher V_{dd}), DIBL, and the onset of parasitic leakage currents at high V_{dd} where the quadratic dependence of gate leakage on V_{dd} becomes relatively higher. Fig. 2.13 shows the modeled (using (29)) and simulated gate delay of a DLS inverter versus V_{dd} showing good agreement at low V_{dd}. At higher supply voltages, V_{ih} can begin to flatten out as shown in Fig. 2.10 which means the gate delay will decrease since Y no longer needs to charge to a higher voltage.

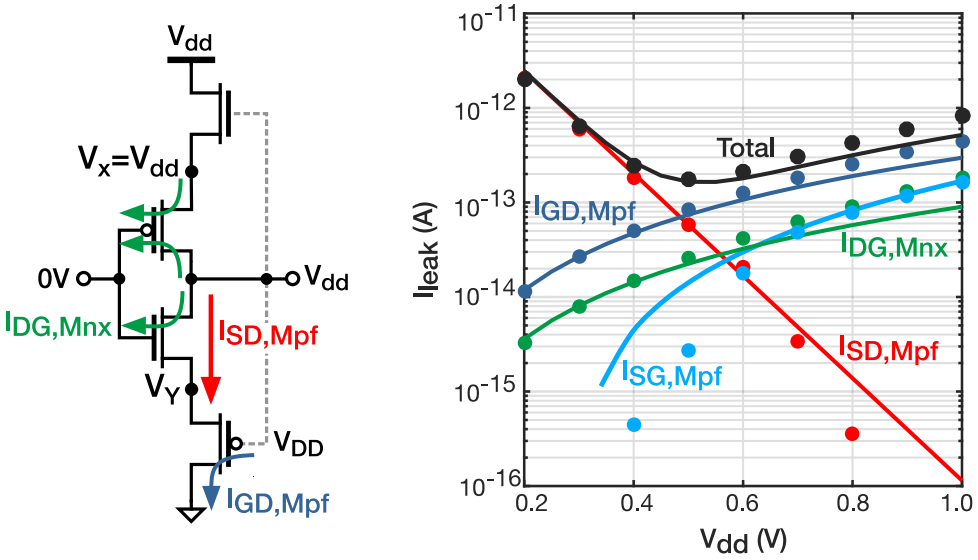

Fig. 2.14. Modeled and simulated leakage current components in a DLS gate with input A=0V versus supply voltage V_{dd}. Models are shown by lines, simulated values are shown by the marker dots.

2.2.3 Power

The leakage power of a DLS gate can be assessed by continuing the previous approach of identifying the most significant leakage currents that are present in each area of the gate. Using the example in Fig. 2.14 when A=0, the output will be charged to V_{dd}. The most obvious source of leakage current is the subthreshold leakage through the PD network, which is represented by the source-drain current of M_{pf}. The voltage V_{dd} present on the output also allows some current to leak through the gates of M_{nx} and M_{px}, as well as through the gate of the footer transistor M_{pf}. Together, these sources lead to a total approximate leakage current:

$$I_{leak} = I_{GD,Mpf} + I_{SD,Mpf} + I_{DG,Mnx} \tag{30}$$

Where $I_{SD,Mpf}$ considers V_y to be in the steady-state voltage determined by (17). Applying the models from (9) and (14), The plot in Fig. 2.14 shows a comparison of the modeled and simulated total leakage I_{leak} from V_{dd}. The results reveal the interesting characteristic of a minimum-leakage point that occurs near the middle of the V_{dd} range. As expected, the increasing V_{dd} causes V_y to increase which further suppresses the subthreshold leakage $I_{SD,Mpf}$, but eventually these savings will be negated by the rising gate leakage from M_{pf} and

M_{nx}. There is good agreement between model and simulation at low V_{dd} where subthreshold currents dominate, while deviations between model and simulation occur at high V_{dd} due to the lack of an exponential scaling term being included in the gate leakage models ((9) is used instead of the full model in (8)). From these results, it is evident that the most important leakage components are $I_{SD,Mpf}$ and $I_{GD,Mpf}$. If we ignore DIBL and the body effect, a new expression can be created to estimate the total leakage power based on these two most significant contributors

$$P_{leak} \approx V_{dd} I_{0,pf} e^{-\frac{V_{dd}+V_{th,nx}+V_{th,pf}}{(n_{nx}+n_{pf})v_T}} + k_{o,pf} V_{dd}^3 \tag{31}$$

More importantly, we can leverage these results to determine the supply voltage at which the total leakage power is minimized. The form of (31) does not allow for easy differentiation, so a usefully simple approach is to solve for the V_{dd} at which $I_{GD,Mpf}$ equals (and subsequently overtakes) $I_{SD,Mpf}$. Keeping DIBL and the body effect temporarily ignored, this yields the approximate optimum supply voltage for power minimization in a DLS logic gate

$$V_{dd,opt} = 2v_T(n_{nx}+n_{pf})W_{-1}\left(\frac{\sqrt{I_{0,pf}}e^{-\frac{V_{th,pf}+V_{th,nx}}{2v_T(n_{nx}+n_{pf})}}}{\sqrt{k_{0,pf}}2v_T(n_{nx}+n_{pf})}\right) \tag{32}$$

Fig. 2.15 shows the simulated P_{leak} of a DLS inverter along with the simulated model from (31), and the identified $V_{dd,opt}$ from (32) is also identified. To further validate the mode, these results are obtained for two different DLS designs, where the design in Fig. 2.15(a) is the one shown previously in Fig. 2.14 where all transistors use similar threshold voltages, while the design in Fig. 2.15(b) chooses M_{nx} and M_{px} to be a higher threshold voltage than M_{nh} and M_{pf} (in accordance with (22)). The difference in transistor threshold voltages directly maps into (2.15), as well as the other models from this section. The derived $V_{dd,opt}$ model matches both cases well, shifting by around 200mV between the two DLS designs and falling reasonably close to the true simulated $V_{dd,opt}$.

The dynamic power of DLS logic can be quite easily modeled by starting with (6), where the only component that needs to be updated is the total gate capacitance C_{gate}. In contrast

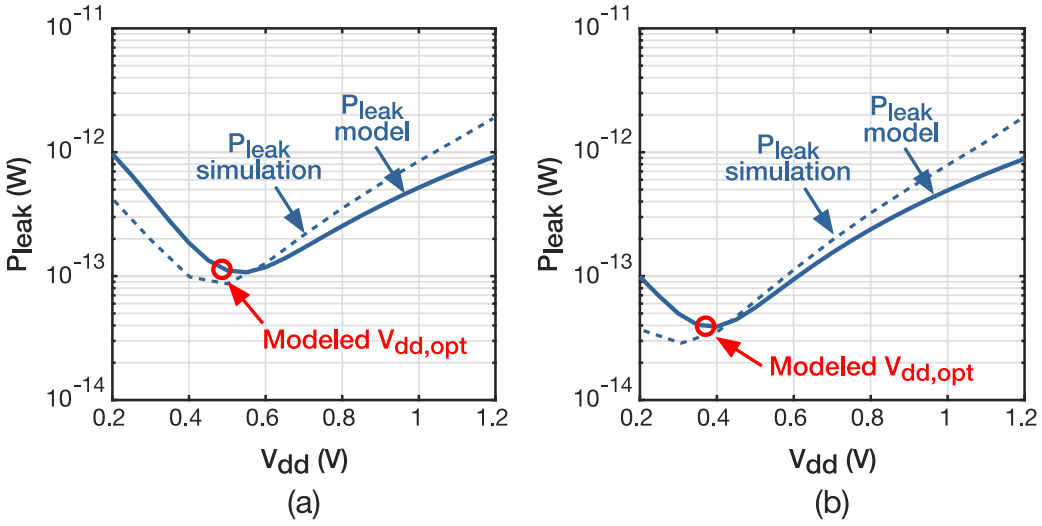

Fig. 2.15. Modeled and simulated leakage power P_{leak} for two different DLS gate designs with the modeled optimum V_{dd} for minimizing leakage power.

with a static-CMOS logic gate, the DLS transistors (M_{nh} and M_{pf}) add an extra amount of capacitance to C_{gate}. Therefore, we can say that

$$C_{gate,dls} = C_{Mnx} + C_{Mpx} + C_{Mnh} + C_{Mpf} \tag{33}$$

and therefore

$$P_{dynamic} = \alpha f_{clk} N C_{gate,dls} V_{dd}^2 \tag{34}$$

At this point, it can be noted that for low leakage, M_{pf} and M_{nh} should clearly have very low subthreshold leakage and a thick gate oxide to keep their gate leakage low. Therefore, techniques such as using RBB, long channel lengths, and HVT devices or thick-oxide I/O devices will help reduce power, but several of these techniques increase the effective C_{gate} which simply increases $P_{dynamic}$. Moreover, leakage reduction in M_{pf} and M_{nh} directly worsens the gate delay t_{gate} in (29) which requires a *high* subthreshold leakage current for high drive strength and quick output transitions. Finally, any design tradeoffs for speed or leakage reduction should still meet the requirements to guarantee positive NMs and SMs based on the guidelines of (22), (24), (26) and (27).

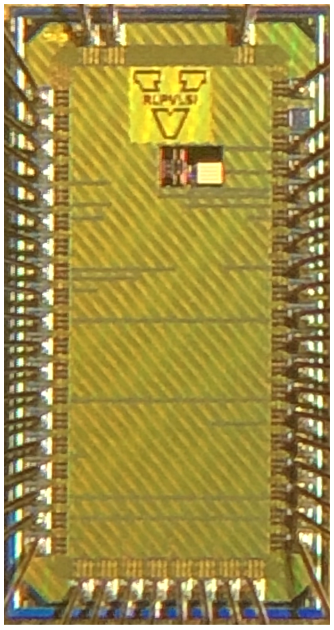

Fig. 2.16. Fabricated DLS test chip in 65-nm CMOS.

2.2.4 Measurement Results

To experimentally assess the performance of DLS logic and verify the prior modeling, a characterization chip was fabricated in a low-power 65nm process. Fig. 2.16, which consumes 2mm^2. Based on the observations from (22), (24), (26) and (27) to guarantee adequate switching margins, we know that the subthreshold leakage of M_{nh} and M_{pf} should be relatively higher than the gate leakage currents (to obtain a large output voltage swing), and that their threshold voltages should be lower than that of M_{nx} and M_{px} (to get input switching thresholds that are not too close to the rails). Additionally, higher subthreshold leakage current from M_{nh} and M_{pf} should enable faster operation, but also result in increased leakage power. Based on these observations, several DLS "flavors" are proposed, which are summarized in Table. 4. The naming convention for each flavor follows a fixed syntax: the first device threshold represents the outer DLS transistors (M_{nh} and M_{pf}), while the second label represents the inner transistors (M_{px} and M_{nx}). For example, the LVT-HVT flavor uses LVT devices for M_{nh} and M_{pf}, while M_{nx} and M_{px} use HVT devices. Different flavors have been designated that make sure the threshold voltages of M_{nh} and M_{pf} are greater than or equal to the threshold voltages of M_{nx} and M_{px}, where as an example the HVT-HVT flavor is expected to achieve

DLS Flavor	M_{nh}/M_{pf}	M_{nx}/M_{px}	Inverter Size
HVT-HVT	HVT[a], W=6×W_{min}	HVT, W=×W_{min}	4.3μm^2
LVT-HVT	LVT[a], W=6×W_{min}	HVT, W=×W_{min}	4.3μm^2
LVT-LVT	LVT[a], W=6×W_{min}	LVT, W=×W_{min}	4.3μm^2
IO-IO	IO, W=×W_{min}	IO, W=×W_{min}	6.78μm^2
IO-IO-FBB	FBB IO[b], W=×W_{min}	IO, W=×W_{min}	6.78μm^2

[a] M_{pf} is forward-body-biased to 0V.
[b] M_{pf} and M_{nh} both forward-body-biased to 0V, V_{dd}, respectively.

Table 4. Summary of DLS flavors in a 65-nm technology that intend to reach different reliability and performance points.

lower power than the LVT-LVT flavor, but also risk the possibility for (26) and (27) to not be met due to the HVT devices having lower subthreshold leakage. Thick-oxide I/O devices are also chosen for two flavors (IO-IO, IO-IO-FBB) to investigate how performance will be affected if gate leakage can effectively be made negligible. Due to the limited availability of discrete threshold voltage selection in these devices (which is dependent on the actual technology), a workaround to decrease the threshold voltage (and increase subthreshold leakage) of M_{nh} and M_{pf} is to Forward Body Bias (FBB) them to the rail voltages (IO-IO-FBB). That is, M_{nh} sits in a p-well biased to V_{dd}, and M_{pf} sits in an n-well biased to 0V. An important note here is that the difference in threshold voltage that are typically present between NMOS and PMOS usually warrants an FBB applied only to the footer M_{pf} transistor, which is easily done since it already sits in its own n-well that can be tied to 0V. Prior DLS implementations have also performed this [1]. For this work, the M_{pf} FBB is performed in all flavors except the IO-IO flavor, which forgoes the M_{pf}-only FBB to enable a better contrast with the fully-FBB IO-IO-FBB style.

The fabricated test chip contains several structures to assess the performance of these DLS flavors. Note that not all of the flavors in Table 4 are implemented for every type of test structure. 29-stage ring oscillators are used to measure the gate delay versus supply voltage. Arrays of 1000 inverters are tied together single V_{dd} pins (one V_{dd} pin per flavor) to measure leakage power. The inputs are all tied together to be able to measure input state-dependent leakage (0V input shown in this dissertation, but the results for A=V_{dd} are not significantly different), and the outputs of all inverters are simply floating. Using an array of parallel 1000 inverters increases the order of magnitude of the measurement to try limitations

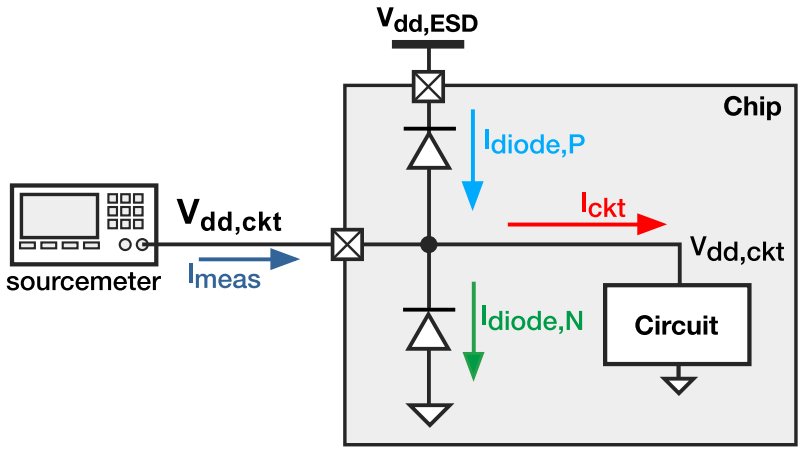

Fig. 2.17. Measured VTCs from 32 LVT-LVT test DLS inverters in a single die operating at V_{dd}=0.25V. Input and output voltages are both normalized to the V_{dd}.

of the testing setup, particularly with respect to accuracy and precision, which will be briefly explained here. Fig. 2.17 demonstrates how the accuracy of the leakage measurement is affected by the parasitic leakage of the Electrostatic Discharge (ESD) protection diodes. The fabricated DLS test chip uses a simple ESD protection scheme with two diodes that will shunt excess current current to ground or the ESD voltage ($V_{dd,ESD}$) if the actual pad voltage gets too high or too low. The pad voltage (in this case, the V_{dd} of the leaking inverter test structures) is always in a range that keeps the diodes turned off, but there is still a reverse bias leakage current through each diode, reflected by (13). Leakage through the PU diode adds current into the measurement node ($I_{diode,P}$) that decreases the measured current (I_{meas}) on the sourcemeter, while the PD diode adds a parasitic leakage ($I_{diode,N}$) that increases the measured current on the sourcemeter. Even with 1000 inverters in parallel, if we assume they at most 1pA of leakage per device, that is only total a measurement current of 1nA which is low enough to be affected by the leakage current of the PU and PD diodes. To further mitigate the impact of diode leakage, the ESD voltage is always kept at twice the pad voltage $V_{dd,ckt}$ to make both diodes see an equal voltage drop such that their leakage will mostly cancel out. Mismatch between the diodes can still skew the measurement (simulation shows there can be a difference of up to several pA), so the measurement can be augmented by subtracting the simulated diode current mismatch. Fig. 2.18 shows the measured leakage and delay of the DLS test structures versus supply voltage for 20 dies (1 test structure, RO or leakage

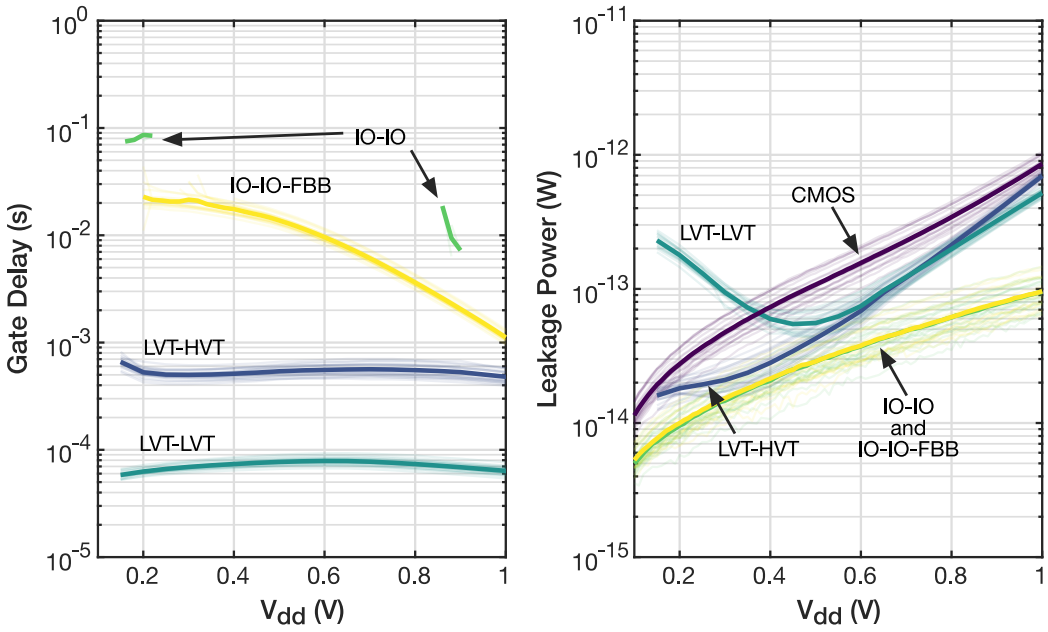

Fig. 2.18. Measured inverter delay and inverter leakage versus supply voltage for several DLS flavors, measured across 20 dies. Leakage for a traditional HVT static-CMOS inverter is also shown. Bolded lines show the average values across all measured dies.

array, per die). Leakage is normalized the amount per single inverter in the leakage array. As expected, the flavors whose M_{nh} and M_{pf} have the highest subthreshold leakage also have the lowest delay, which was predicted by (29). The IO-IO flavor did not function across the full range, with only a few of the tested chips operating sporadically at different supply voltages. The LVT-LVT flavor has the highest relative subthreshold leakage to gate leakage ratio, giving rise to the minimum leakage point that was demonstrated in Fig. 2.14. Switching the internal transistors to HVT devices, the LVT-HVT flavor reduces the subthreshold current through the leaking PD network, causing gate leakage to dominate across more of the supply voltage range (again, refer to Fig. 2.14). The IO-IO and IO-IO-FBB flavors have the lowest leakage as expected due to their thick oxides suppressing gate leakage. While they achieve less leakage power than the traditional static-CMOS structures, it is expected that the parasitic leakage interfered with this measurement due to the true leakage power of these flavors being masked by the leakage floor of the ESD diodes. To measure the VTC of a DLS inverter, the structure in Fig. 2.19(a) is used to buffer the inverter voltage out with a unity-gain analog buffer (with thick-oxide devices) that protects the output voltage of the DLS gate from parasitic loading/leakage. 32 VTC measurement structures are used per

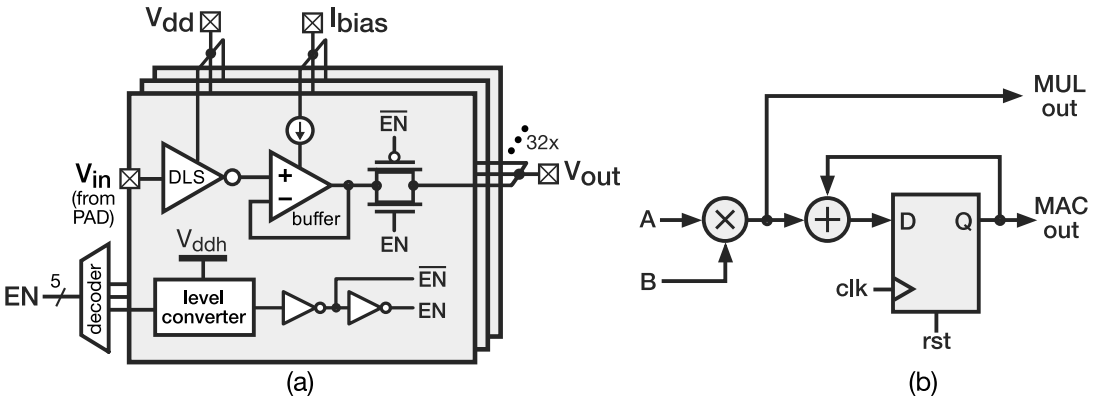

Fig. 2.19. Circuit setup for measurement of DLS inverter VTCs (a) and functional diagram of the implemented Multiply-Accumulate (MAC) circuit.

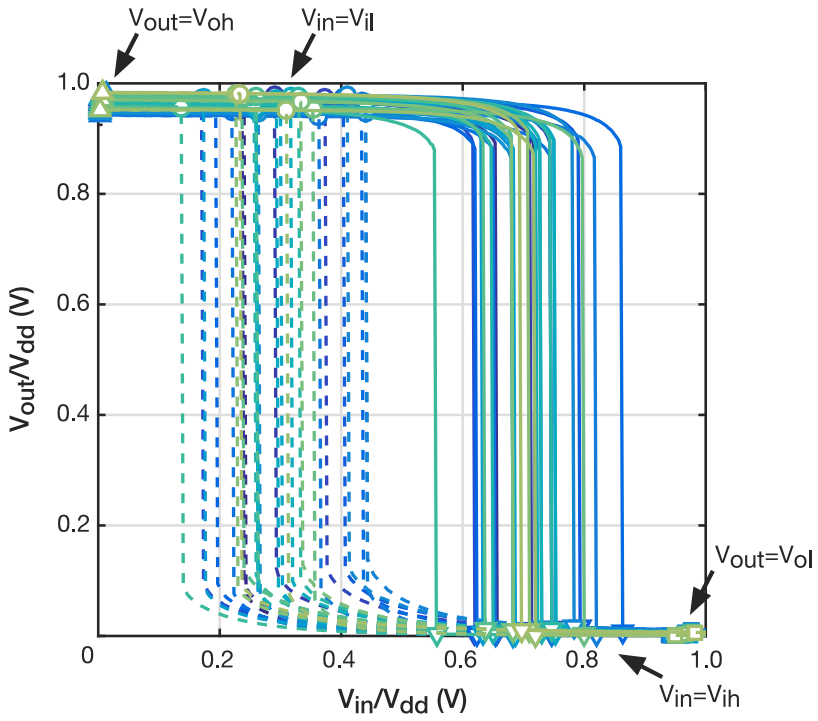

Fig. 2.20. Measured VTCs from 32 LVT-LVT test DLS inverters in a single die operating at V_{dd}=0.25V. Input and output voltages are both normalized to the V_{dd}.

flavor to help assess within-die variation, so to enable their outputs to be multiplexed (so they can be individually-measured from the same chip output pad) a transmission gate is used as an analog switch, and it is designed with thick-oxide I/O devices to avoid gate leakage and the gates are overdriven with a local level converter to ensure low on-resistance and low off-state leakage. The unity gain amplifier bias current is fully tunable, and this setup also enables the offset voltage of every amplifier to be sampled prior to measuring a VTC

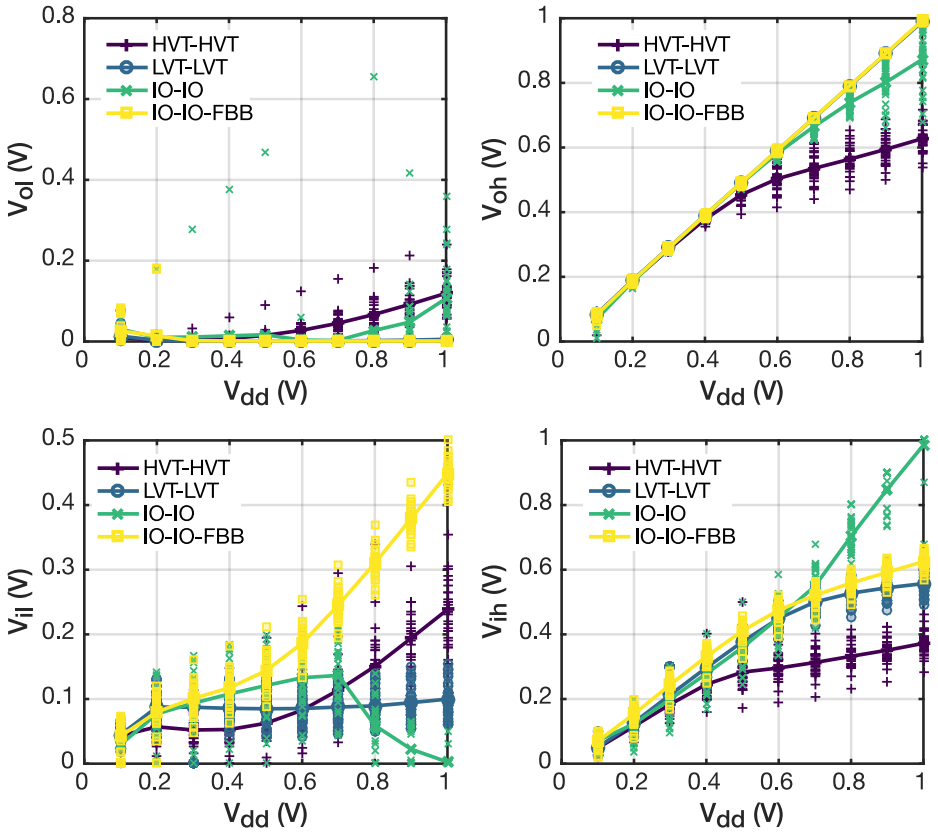

Fig. 2.21. Measured DLS inverter DC input and output characteristics versus supply voltage, obtained from all 32 measured VTCs for each of the DLS flavors in one die.

curve to obtain a more accurate measurement. Fig. 2.20 shows the 32 measured VTCs of the LVT-LVT flavor from one die while operating at a 0.25V V_{dd}. Markers are shown for the identified V_{oh}, V_{ol}, V_{ih}, and V_{il}, highlighting a notable spread in V_{il} and V_{ih} amongst the 32 measured inverters. The algorithms to obtain these data points are given in Appendix C. The measurement in Fig. 2.20 was performed across all supply voltages and all DLS flavors for a single die. The resulting DC characteristics (V_{oh}, V_{ol}, V_{ih}, and V_{il}) for the HVT-HVT, LVT-LVT, IO-IO, and IO-IO-FBB are shown across V_{dd} in Fig. 2.21. As predicted by (26) and (26), the flavors with the highest threshold voltages (lowest subthreshold leakage currents) for M_{nh} and M_{pf} have the worst output swing. That is, the HVT-HVT and IO-IO devices have the highest V_{ol} and lowest V_{oh}. The predictions for input switching threshold are also reasonably accurate – only the IO-IO-FBB flavor has a large skew in threshold voltage between the internal and header/footer transistors (due to the FBB of the header/footer lowering their threshold voltage), so its V_{il} is the highest, which is reflected by (24). Since the other flavors

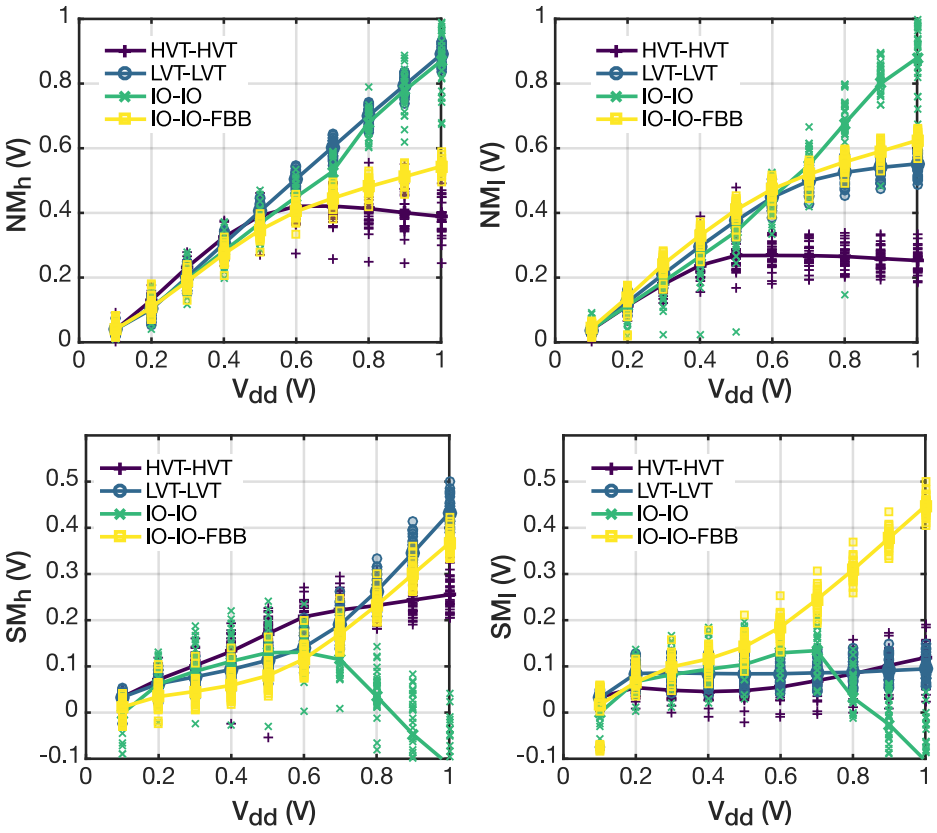

Fig. 2.22. Measured DLS inverter operating margins versus supply voltage, obtained from all 32 measured VTCs for each of the DLS flavors in one die.

have a smaller relative threshold gap between their $V_{th,nh}$ and $V_{th,px}$, their V_{il} hovers more closely to 0V and is dictated more by the secondary effects of DIBL and the body effect as shown in (23). The IO-IO flavor is the only one without a FBB M_{pf}, leading its $V_{th,nx}$ and $V_{th,pf}$ to be the closest out of all the flavors and a V_{ih} close to V_{dd} as predicted by (24). Flavors such as the HVT-HVT that are less dominated by subthreshold current see a low V_{ih} that flattens out at high V_{dd} as shown in Fig. 2.10. The results from Fig. 2.21 are now transformed into the operating margins NM and SM defined by Table 3. The results are shown across V_{dd} in Fig. 2.22. As expected, most of the flavors maintain a NM above the traditional $V_{dd}/2$ limit, although this deteriorates at high V_{dd} for the flavors that begin to see a reduced output swing (HVT-HVT and IO-IO in particular). There a few isolated cases of very low NMs, but none of the measured inverters (in this single chip) demonstrated a negative NM at any V_{dd}. Conversely, the SMs are mostly limited below $V_{dd}/2$, with a few individual inverters from all flavors showing negative NMs at different supply voltages, which would indicate functional

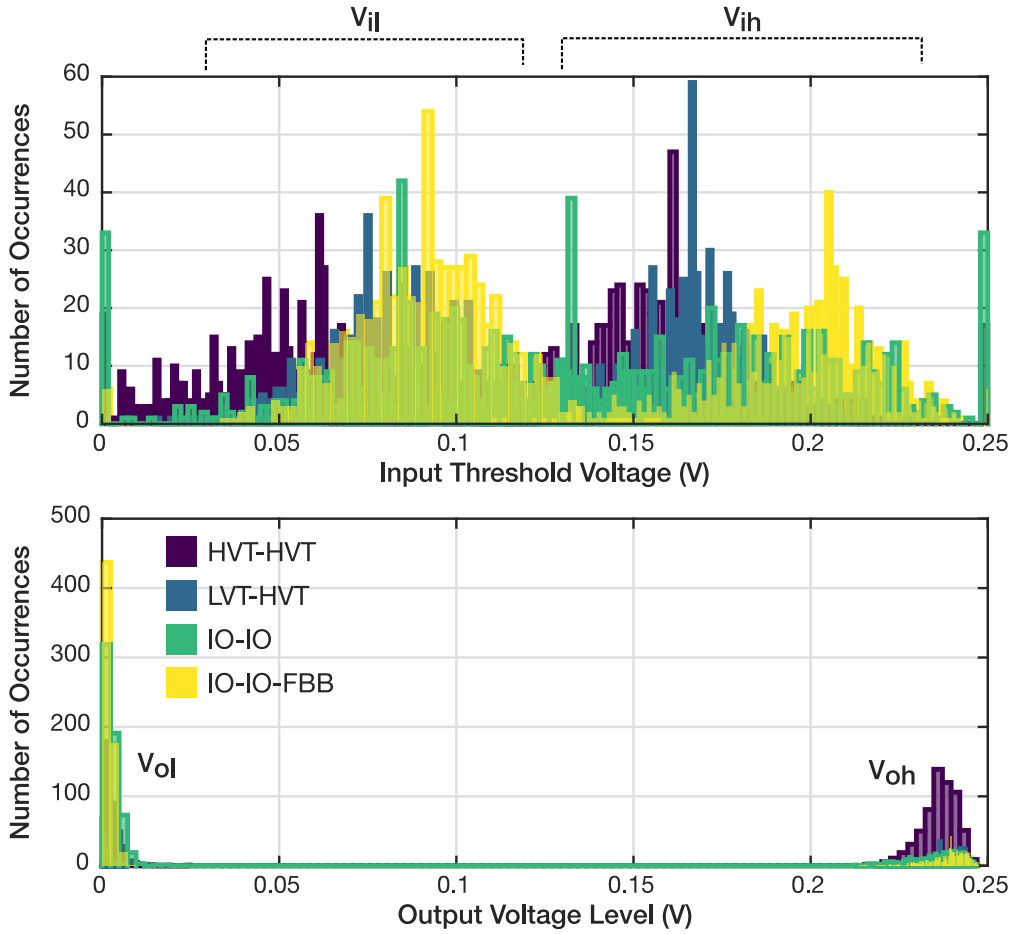

Fig. 2.23. Measured distribution of input threshold voltages (V_{il} and V_{ih}) and output voltage levels (V_{ol} and V_{oh}) of all DLS structures and all chips at a V_{dd} of 0.25V. With 32 measured VTCs from each of 20 chips, this yields 640 measured points for each structure type.

failure.

To better assess the yield of the different flavors, a fixed supply voltage of V_{dd}=0.25V is chosen, and the VTCs of 4 flavors (HVT-HVT, LVT-HVT, IO-IO, and IO-IO-FBB) are measured for every chip. With 32 inverters per chip and 20 total chips, this provides 640 measured VTCs per flavor. Fig. 2.23 shows a histogram of the combined (across all chips) DC characteristics (V_{il}, V_{ih}, V_{ol}, and V_{oh}) for the each flavor. Variation between same-flavor devices that causes their V_{il} and V_{ih} distributions to overlap (for example, one inverter's V_{il} is higher than another's V_{ih}) is not an issue because of the hysteresis effect, however the V_{il} and V_{ih} distributions should fit within the interior limits of the V_{ol} and V_{oh} distributions. This will be discussed in more detail shortly. Table 5 breaks down the data from Fig. 2.23 to summarize the the statistical properties of the DC characteristics across all 20 measured chips. For each

DLS Flavor	Within-die μ (mV)	Within-die σ (mV)	Die-to-die σ (mV)
V_{il}			
HVT-HVT	51	22	4.3
LVT-LVT	84	21	5.1
IO-IO	91	36	5.9
IO-IO-FBB	90	21	3.6
V_{ih}			
HVT-HVT	153	28	4.8
LVT-LVT	167	16	3.0
IO-IO	170	42	11
IO-IO-FBB	195	22	5.0
V_{ol}			
HVT-HVT	2.6	2.7	0.74
LVT-LVT	1.4	1.0	0.3
IO-IO	1.4	45	7.5
IO-IO-FBB	4.3	11	5.7
V_{oh}			
HVT-HVT	231	22	6.5
LVT-LVT	238	4.1	1.1
IO-IO	236	5.6	2.3
IO-IO-FBB	238	4.8	0.8

Table 5. Measurement of variation in the DC input and output characteristics of the DLS inverters from 20 chips operating at a 0.25V V_{dd}.

flavor listed, the within-die average is the mean of all the individual within-die averages, the within-die σ is the mean of all the individual within-die σ values, and the die-to-die σ is the σ of all the individual within-die averages. For example, for any flavor, if the 32 VTCs from chip *i* have values with mean of μ_i and standard deviation σ_i, then the listed within-die mean is $\mu_{WID} = \text{mean}(\mu_{i=1}, \mu_{i=2}, \ldots \mu_{i=N})$, the within-die σ is $\sigma_{WID} = \text{mean}(\sigma_{i=1}, \sigma_{i=2}, \ldots \sigma_{i=N})$, and the die-to-die σ is $\sigma_{D2D} = \sigma(\mu_{i=1}, \mu_{i=2}, \ldots \mu_{i=N})$ for N=20 chips. Table 6 shows the measured statistics of the operating margins. computed from the same measurements used in Table 5, of all the flavors across all 20 chips. The within-die and die-to-die measurements are calculated as previously mentioned. The individual yield percentage reflects the number of inverters (out of the 640 total across all chips for each flavor) that pass the yield test (NM_h, NM_l, SM_h, and SM_l all must be positive) based on just their own individual characteristics. For example, any inverters SM_h is computed with its own V_{oh} and V_{ih} values. This individual yield result is helpful for confirming the possibility of block-level *failure* by identifying if gate-level failures are possible (if a gate fails to operate with copies of itself, it has a reasonable chance to fail at driving whatever gate(s) it connects to in a real block). However, it does not offer confidence in block-level *success*, which would require certainty that any two random gates

DLS Flavor	Within-die μ (mV)	Within-die σ (mV)	Die-to-die σ (mV)	Individual Yield (%)	Die Yield (%)
NM_l					
HVT-HVT	150	27	4.9	100%	100%
LVT-LVT	166	15	2.9	100%	100%
IO-IO	156	48	10	100%	15%
IO-IO-FBB	191	22	65	99.84%	80%
NM_h					
HVT-HVT	180	30	5.6	100%	60%
LVT-LVT	154	21	4.7	100%	100%
IO-IO	144	33	5.0	100%	100%
IO-IO-FBB	148	21	3.4	99.84%	95%
SM_l					
HVT-HVT	48	23	4.6	97.18%	10%
LVT-LVT	83	20	5.0	100%	100%
IO-IO	71	72	10	94.84%	15%
IO-IO-FBB	86	29	7.5	98.9%	80%
SM_h					
HVT-HVT	78	44	9.8	97.34%	50%
LVT-LVT	70	16	2.7	100%	100%
IO-IO	66	41	9.6	94.8%	5%
IO-IO-FBB	43	22	4.8	98.9%	15%

Table 6. Measurement of variation and yield in the operating margins (SM and NM) of the DLS inverters from 20 chips operating at a 0.25V V_{dd}.

could work together if their characteristics were individually-skewed enough to result in a joint worst-case scenario. Therefore, rather than calculating NMs and SM based on a gate's own input and output characteristics, a conservative approach to estimate yield is to assert that positive NMs and SMs should exist between the worst-case conditions of any two individual inverters. For example, the entire die can be assumed to have positive SM_h if the lowest V_{oh} out of all the inverters in the die exceeds the highest V_{ih} out of all the inverters in the die. Of course, this approach is conservative because it is not guaranteed that two incompatible gates will actually be connected together. This resulting calculation is named as the "Die Yield" in Table 6 and represents the number of dies (out of the full 20) that would still work (positive NMs and SMs if their own two worst-case inverters happened to be connected together. Two important notes here are first that this yield is only taken from 32 inverters, and would be significantly lower if hundreds or thousands of gates were considered. Second, these measurements were taken at a V_{dd} of 0.25V where the flavors from just one die show a visible degradation in their operating margins (Fig. 2.22). If V_{dd} were increased, it is expected that these yield numbers would rise, since the operating margins are much more positive at higher V_{dd}.

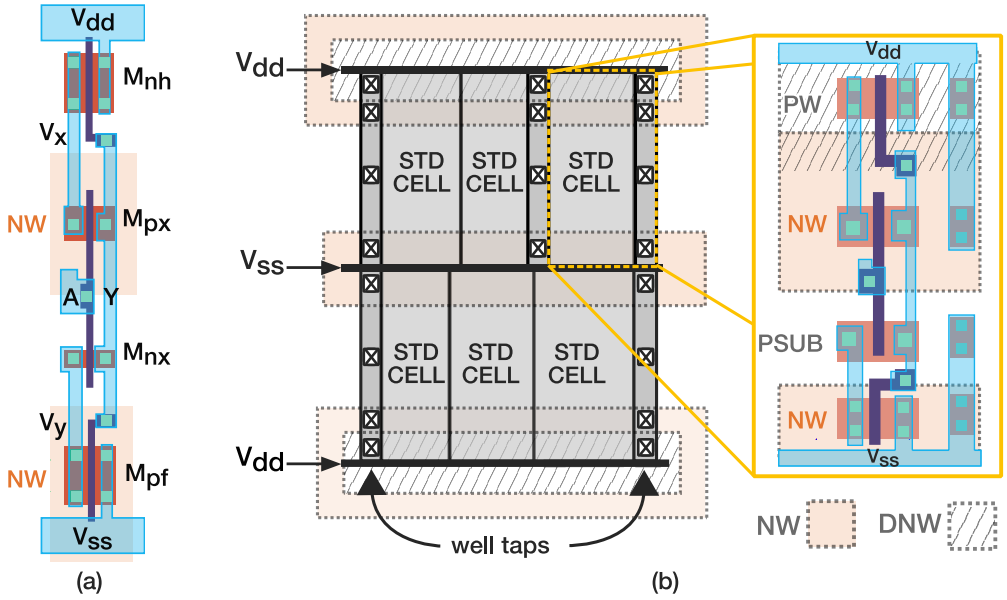

Fig. 2.24. Layout implementation of a DLS inverter (a) and standard cell array implementation for Forward Body Bias (FBB) IO-IO DLS logic flavor (b).

To demonstrate DLS logic working in an actual circuit, an 8-bit Multiply-Accumulate (MAC) circuit is used, whose architecture is shown in Fig. 2.19(b). Two 8-bit inputs, A and B, are multiplied with a purely-combinatorial multiplier, and the multiplied readout (which is 16 bits) is directly readable. An accumulator is used to add the multiplied result to a value held in a data register. The multiplier and adder will work asynchronously, but a positive edge on the MAC clock will latch the adder output into the register, which will instantly begin a new addition operation. A standard cell library of DLS gates is created for each flavor that requires only an inverter, a digital buffer, a 2-input NAND, a 2-input NOR, and a D flip-flop with asynchronous reset. These are the only gates needed since the MAC circuit synthesized from a full commercial standard cell library for static-CMOS logic uses the same gate selection, leading to an identical gate netlist. To avoid the additional design effort of transmission gate-based flip-flops (interactions between DLS gates and standard transmission gates are difficult problem to deal with, as will be shown in Section 3.2.1), the DLS D flip-flop is simply constructed with NAND and NOR gates. Fig. 2.24(a) shows an example layout implementation of a DLS inverter, where the spatial arrangement of the gates actually matches the schematic representation with M_{nh} sitting on top of M_{px} and M_{pf} sitting below M_{nx}. When standard cells are then combined along a track, the nwells and sections of

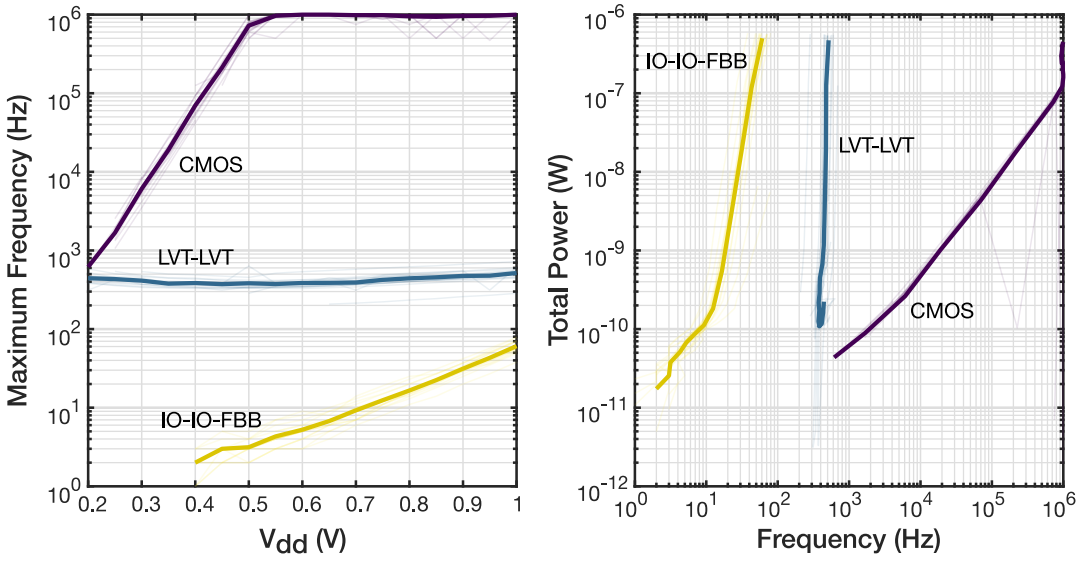

Fig. 2.25. Measurements of the DLS MAC circuits with comparison to HVT CMOS. The maximum operating frequency versus supply voltage is shown, as well as a pareto-optimal plot for power versus energy.

p-type substrate can then easily be biased with well tap cells. Fig. 2.24(b) shows how the IO-IO-FBB cells are implemented into a full array, with an additional section of deep n-well inserted into each cell layout beneath M_{nh}. When cells are tiled together onto mirrored tracks, an isolated p-well is created which can then be tied off with end cap cells. The well tap cells can then easily bias the M_{nh} p-well body to V_{dd} and the M_{pf} n-well body to V_{ss}. The standard cells for each DLS flavor are characterized for timing using a standard commercial tools and the MAC circuit for each DLS flavor is fully synthesized, placed, and routed. The MAC circuits were tested by first multiplying two inputs together and then sending a burst of clock pulses in at a fixed frequency, and then checking the resulting MAC value after the burst is completed to verify if its value is correct. The total power is also measured while the clock pulse burst is active. The results are shown in Fig. 2.25, which show that the CMOS maximum frequency increases exponentially with V_{dd}, while the LVT-LVT maximum frequency remains nearly constant. While the LVT-LVT flavor does reach much higher frequencies than prior implementations of DLS logic and showed full yield across the supply voltage range for all 20 dies (matching the predictions from Table 6), it consumes just as much power as the static-CMOS MAC due it having comparable leakage current (which will be shown next) and a higher effective gate capacitance that increases dynamic power, which was modeled by (34). The IO-IO-FBB maximum frequency increases exponentially with V_{dd} but none of the

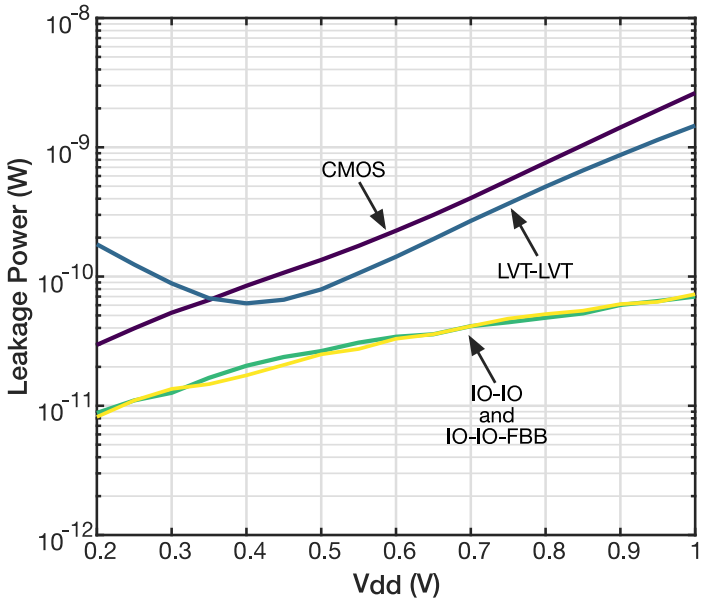

Fig. 2.26. Measured leakage power versus supply voltage of the DLS and static-CMOS MAC circuits.

dies were able operate below 0.4V. The IO-IO MAC did not work at all, which correlates with the yield predictions from Table 6. A pareto-optimal plot is shown for power versus operating frequency, showing that the static-CMOS MAC offers the lowest power until frequencies under 10Hz are desired, at which point the IO-IO-FBB design offers less total power consumption. As mentioned before, it is expected that parasitic leakages in the measurement setup skew these results in favor of the static-CMOS logic. The measured leakage power versus supply voltage for each MAC circuit is shown in Fig. 2.26. These results correlate well with the individual inverter leakage measurements from Fig. 2.18, showing that leakage of the LVT-LVT flavor increases at low V_{dd} and that the thick-oxide flavors achieve the lowest leakage across the V_{dd} range.

2.2.5 Conclusion

This section analyzed the DC and transient performance of DLS logic and derived theoretical models for key performance metrics such as input threshold voltages, leakage power, and gate delay. The models are used to develop several key design considerations for DLS logic which are then used to develop multiple “flavors” that seek to meet different performance and reliability targets. A test chip is fabricated in 65-nm CMOS that contains measurement test structures for the developed DLS flavors, and measurement results generally support

expectations based on the derived performance models.

2.3 Performance Scaling for DLS Logic

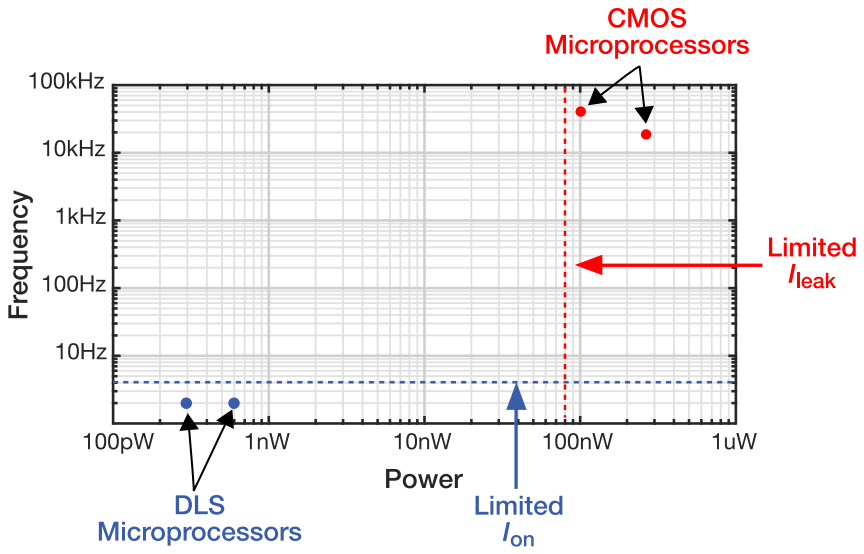

Fig. 2.27. Design space for low-power and low-frequency microprocessors. DLS-based works are limited by their maximum speed [1, 2], while traditional CMOS-based works are limited by their leakage currents [3].

Prior work on DLS logic has experimentally confirmed the leakage reduction benefits, but also shown that DLS is limited to maximum operating frequency well below 100Hz if it is to be designed in a way that achieves less leakage than static-CMOS logic. When compared with traditional static CMOS-based processor designs, a large performance gap exists in the design space that can't be reached by either DLS logic or static CMOS logic, as shown in Fig. 2.27. The dramatic cost for DLS logic to trade off between leakage and delay and reduces its usefulness to systems and applications that can't handle such a large performance trade-off. While Section 2.2 of this dissertation focuses on improving the native performance flexibility of DLS logic, this section focuses on physical modifications to the gate design to improve the performance flexibility. The contribution of this section is a Scalable Dynamic Leakage Suppression (SDLS) variant and associated hardware for enabling a performance scalable DLS implementation that can smoothen the leakage-delay trade-off between DLS and static-CMOS logic styles. The reference publication for the material in this section is [DT17].

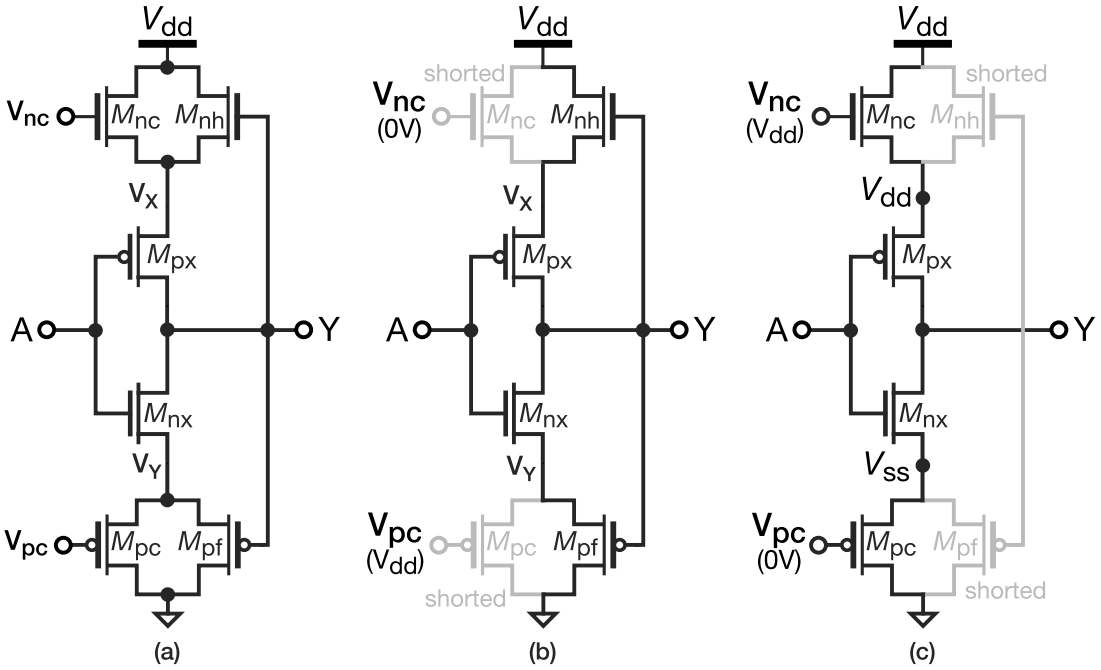

Fig. 2.28. (a) Design of Scalable Dynamic Leakage Suppression (SDLS) logic gate, (b) DLS-based operation of SDLS gate when the control voltage is low, and (c) CMOS-based operation of SDLS gate when control voltage is high.

2.3.1 SDLS Design

Possibly the strongest and most common knob for adjusting the leakage and delay of traditional digital logic is the supply voltage V_{dd}. DLS logic is driven by leakage currents which only have a small dependence on V_{dd} due to Drain-Induced Barrier Lowering (DIBL), which leads to a poor performance scaling ability from V_{dd} scaling. The SDLS logic style, shown in Fig. 2.28(a), implements a new voltage-scaling approach in which two complementary control voltages, V_{nc} and V_{pc}, are used to transition the gate across a continuous range between DLS and CMOS operating regimes, trading off between leakage power and gate delay. Compared to traditional DVFS, this constant-V_{dd} approach eliminates the need for level shifters and avoids regulator efficiency losses caused by varying supply levels. In the SDLS logic gate, transistors M_{nc} and M_{pc} are added in parallel to the traditional DLS transistors M_{nh} and M_{pf}, similar to the approach in [2]. If the control voltages V_{nc} and V_{pc} are asserted such that M_{nc} and M_{pc} are disabled, then the SDLS gate will operate like a standard DLS gate, shown in Fig. 2.28(b). Alternatively, if V_{nc} and V_{pc} are chosen such M_{nc} and M_{pc} are active, then the original DLS transistors M_{nh} and M_{pf} will be shorted out and the

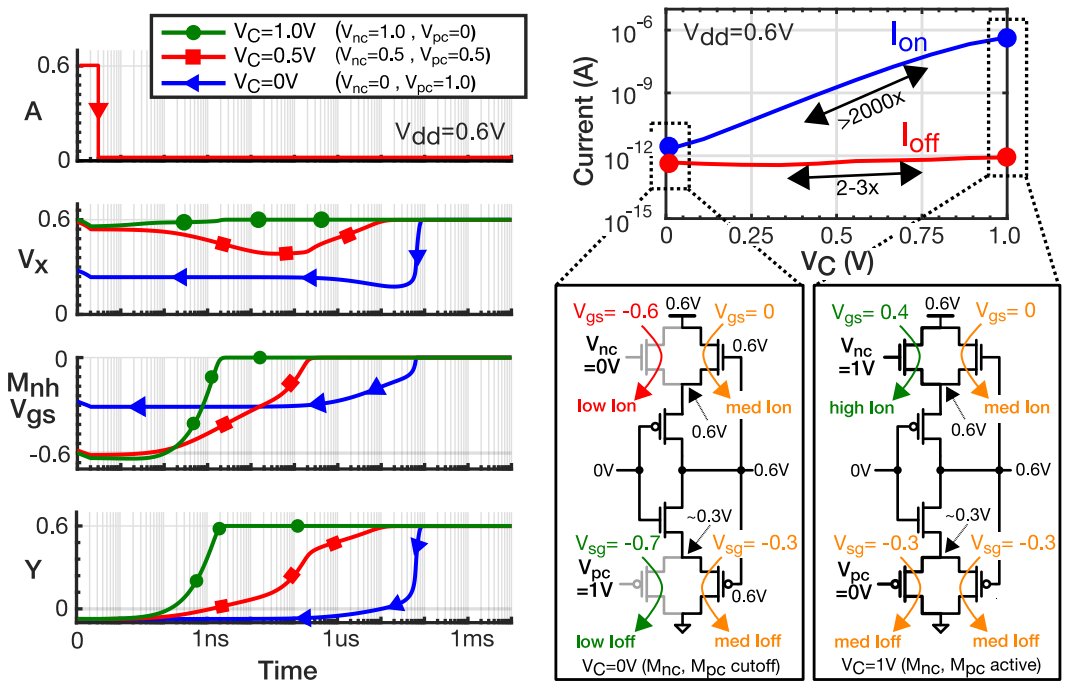

Fig. 2.29. Illustration of how the control voltage affects the operating parameters of the devices in the SDLS gate. Scaling the control voltage up (higher V_c means higher V_{nc} and lower V_{pc}) gives the gate more drive current via M_{nc} and M_{pc}, but also increases their leakage current as well.

internal nodes V_x and V_y will be equal to the V_{dd} and V_{ss}, respectively, such that the SDLS gate will operate like a traditional static-CMOS gate as shown in Fig. 2.28(c). In addition to the proposed control voltage approach for modulating gate performance, two other new design modifications are added here to improve performance scalability and reduce gate area. First, the body of internal PMOS transistors (M_{px}) are tied to V_{dd}, rather than V_x as in [1], in order to save area by sharing internal n-wells with negligible performance degradation. Second, LVT devices are used for the external transistors M_{nh}, M_{nc}, M_{pf}, and M_{pc}, allowing increased sensitivity to V_{nc} and V_{pc} for larger performance tuning range while simultaneously decreasing the large transistor sizes required in [1] and [2]. This is identical to the LVT-HVT DLS flavor shown in Table 4. Fig. 2.29 demonstrates the transient operation of an SDLS inverter during a falling input transition. In the steady state when the input A is high and V_c is 0V (V_{nc}=0V and V_{pc}=1V), the internal node V_x settles to an intermediate voltage (see 2.2.1), pushing M_{nc}, M_{nh}, and M_{px} into super-cutoff mode (negative V_{gs}), reducing leakage through the pull-up network. When A transitions low, M_{px} turns on, allowing Y and V_x to converge. This creates positive feedback by increasing V_{gs} of M_{nh}, allowing more current

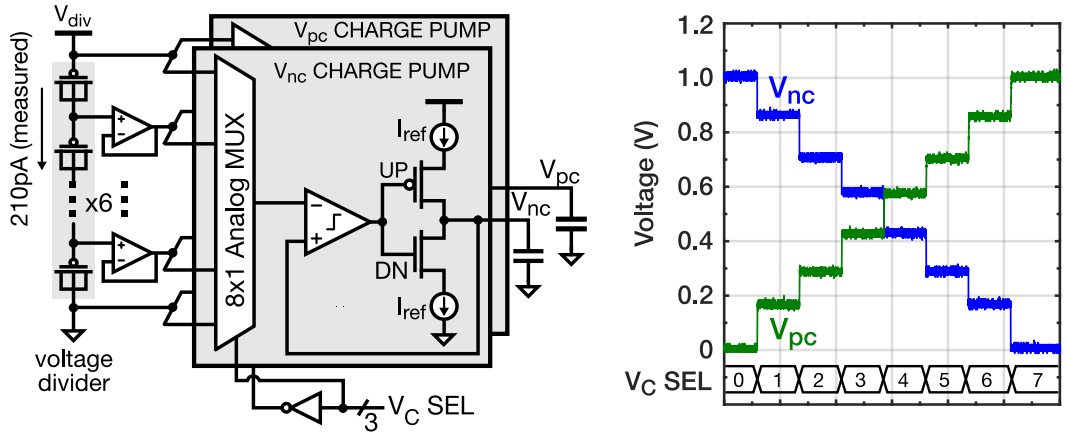

Fig. 2.30. Schematic diagram of VSC and measured oscilloscope plot of the VSC outputs while the control value V_c *SEL* is swept from its minimum to its maximum value. $V_{analog} = 1.2V$, $V_{div} = 1.0V$.

to leak through the pull-up network, further charging Y until it has fully transitioned. If V_c is increased (V_{nc}>0V and V_{pc}<1V), transistor M_{nc} pulls the steady state intermediate voltage of V_x closer to V_{dd}. While this weakens the super-cutoff effect in the pull-up network causing increased leakage, it also accelerates the convergence of Y and V_x, allowing a quicker feedback response from M_{nh} that results in a shorter gate delay. In addition to increasing M_{nh}'s feedback response, M_{nc} provides increased on-current throughout the transition that further improves speed. The same functionality applies during a rising-edge input transition, with V_{pc} tuning the steady-state voltage of V_y.

2.3.2 SDLS Supporting Hardware

To enable control over the SDLS bias transistors, a VSC is designed that uses a digitally-controllable voltage charge pump to adjust the biasing point of the SDLS logic on the fly, similar to the way that DVFS in traditional static-CMOS logic adjusts the supply voltage on the fly. The voltage scaling controller, shown in Fig. 2.30, operates from its own analog-domain supply voltage $V_{analog} = 1.2V$ and generates V_{nc} and V_{pc} (each up to a maximum voltage $V_{div} \leq V_{analog}$) by selecting two complementary reference voltages from a gate leakage-based resistive voltage divider using a signal named V_c *SEL* and then driving the V_{nc} and V_{pc} nodes to the selected voltages with tunable bang-bang pA-level switched current sources (I_{ref}). The tunable pA-level current source uses the design from Fig. 4.8 in Section 4.2. The gate-leakage based resistive divider offers an area efficient solution to voltage division that

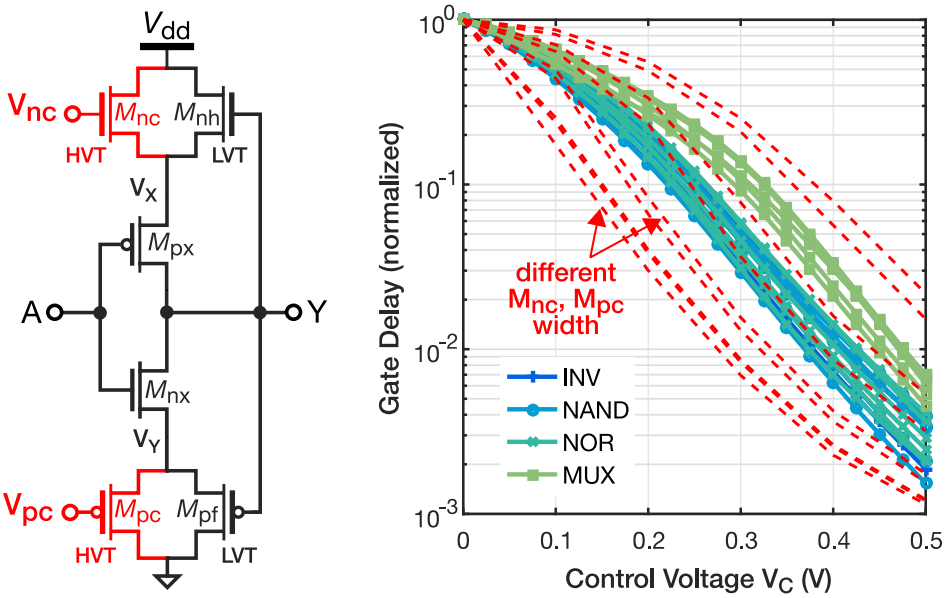

Fig. 2.31. Design of SDLS inverter to be used as a clock delay cell for critical path replication. Changing the widths of M_{nc} and M_{pc} allows the clock cell to achieve a different delay scaling across V_c that accommodates the different intrinsic delay characteristics of SDLS standard cells.

consumes very little quiescent current (measured at 210pA). At the nominal core V_{dd} of 0.6V, a V_{div} of 1.0V is used to overdrive V_{nc} and V_{pc} for a stronger cutoff effect. Fig. 2.30 also shows the measured V_{nc} and V_{pc} waveforms obtained from sweeping V_c *SEL* across its full range. This approach can support a wide range of transition times (between two different V_c *SEL* values) by simply adjusting the value of I_{ref}. To create an adaptive clock generator for the SDLS core, it is necessary to design a circuit that will track the critical path delay of the core. While it is possible to do this with advanced methods that use feedback for in-situ timing error detection for predictive timing errors [58], a more simple approach is to create a replica circuit that replicates the delay of the critical path and use it to clock the core. For systems that perform traditional V_{dd} scaling via DVFS, the key requirement for this replica circuit is that it can capture way that delay scales across V_{dd} for any gate in the standard cell library since each gate can have its own unique delay scaling characteristic. For example, if the critical path is dominated by NAND gates, the delay versus V_{dd} characteristic would be different from a case where the critical path is dominated by AOI gates. In this case for the SDLS core, it is necessary for the critical path replicator circuit to track any combination of standard cells (in series, as a critical path) across the V_c range. Fig. 2.31 shows the normalized delay

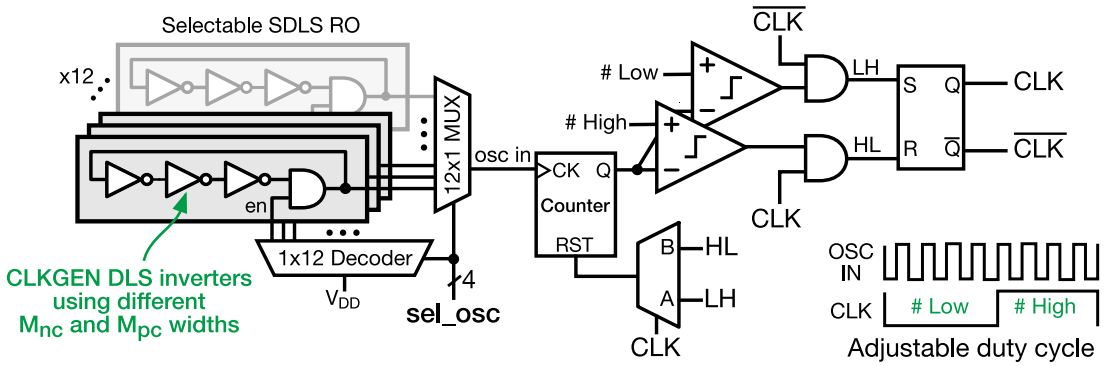

Fig. 2.32. Schematic diagram of the SDLS ACG. A 3-stage SDLS-based ring oscillators is divided down in frequency by a 2-phase programmable frequency divider that allows for customizable frequency and duty-cycle.

versus V_c for several SDLS gates. Note here that $V_c = V_{nc} = V_{dd} - V_{pc}$. To meet any delay characteristic within this range (with optional additional timing margin for safety), an SDLS clock cell inverter is designed where M_{nc} and M_{pc} are implemented with HVT devices to alter the dependence of their current on V_c in contrast with the existing M_{nh} and M_{pf} which are implemented with LVT devices. Then, if the widths of M_{nc} and M_{pc} are tuned, the delay sensitivity to V_c can be strengthened or weakened. Shown by the dashed lines in Fig. 2.31, a range of widths are chosen that allow multiple SDLS clock cells to be created with unique delay characteristics that collectively envelop the characteristics of the native SDLS standard cells. Thus, depending on the exact delay characteristic of the critical path, a ring oscillator constructed with the correct SDLS clock cell inverters will be able to match the characteristic with any desired amount of timing margin. The full adaptive clock generator, shown in Fig. 2.32, uses the SDLS clock cell inverters to replicate the critical path delay of the SDLS core and track the voltage ripple on V_{nc} and V_{pc} to maintain maximum operating frequency of the core. The replica path incorporates two main tuning methods that allow for variable tracking accuracy depending on the desired amount of tuning and calibration. First, the period and duty cycle of the output clock signal are tunable by counting a programmable number of cycles of a 3-stage SDLS clock cell Ring Oscillator (RO) that runs from the same V_{dd}, V_{nc}, and V_{pc} as the processor core. The number of counted cycles is separately tunable for both the high and low levels of the clock output. Second, the SDLS clock cell ring oscillator is selected from a bank of 12 possible ring oscillators to provide selectivity over the delay sensitivity to V_c to help ensure a close match between the replica path delay and the actual

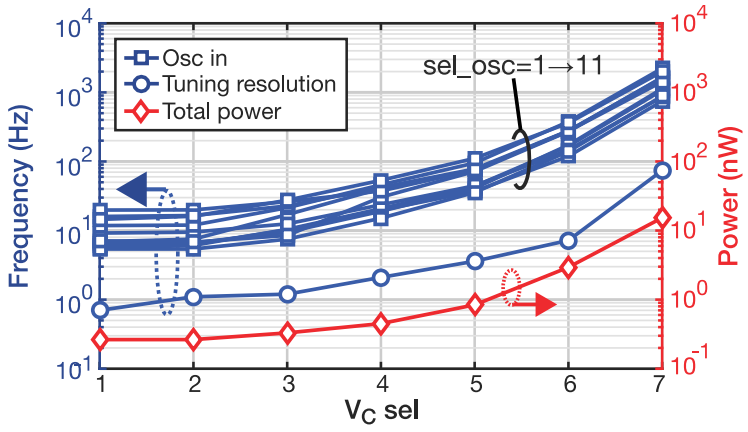

Fig. 2.33. Measured frequency and power of the Adaptive Clock Generator (ACG). Frequencies of each individual SDLS RO are recorded, and the frequency tuning resolution is measured by adjusting the frequency divider value.

critical path delay across V_c mode and V_c ripple. Each 3-stage RO within the bank of 12 ROs uses a different width configuration for M_{nc} and M_{pc} in all of its inverters. Since all of these parameters can be set and changed from memory-mapped registers, the critical path delay can be calibrated at any number of V_c modes, with near-optimum tracking being achieved by tuning the replicator at each V_c mode for a given program. Fig. 2.33 Shows the measurements for each oscillator cell (sweeping sel_osc), the minimum frequency tuning resolution, and the power of the ACG. Besides the 3-stage SDLS RO cells, the remainder of the clock generator logic is implemented with standard CMOS gates to reduce delay overhead.

2.3.3 SDLS RISC-V Core

Figure 2.34 shows the system architecture. The SDLS processor implements a BottleRocket RISC-V core (RV32IMC, derived from Rocket processor [59]) which interfaces to a custom 8KB 6T SRAM macro and an uncore domain that includes system interconnect, SPI master and GPIO peripherals, and a memory-mapped DVFS control register that allows the core to tune the VSC control voltage mode as well as the clock generation settings for the adaptive clock generator. The core and uncore logic is synthesized from an SDLS standard cell library with a fully-automated place and route, and use static CMOS-style clock buffers in order to minimize insertion delay and improve slew rate relative to SDLS inverters. The library

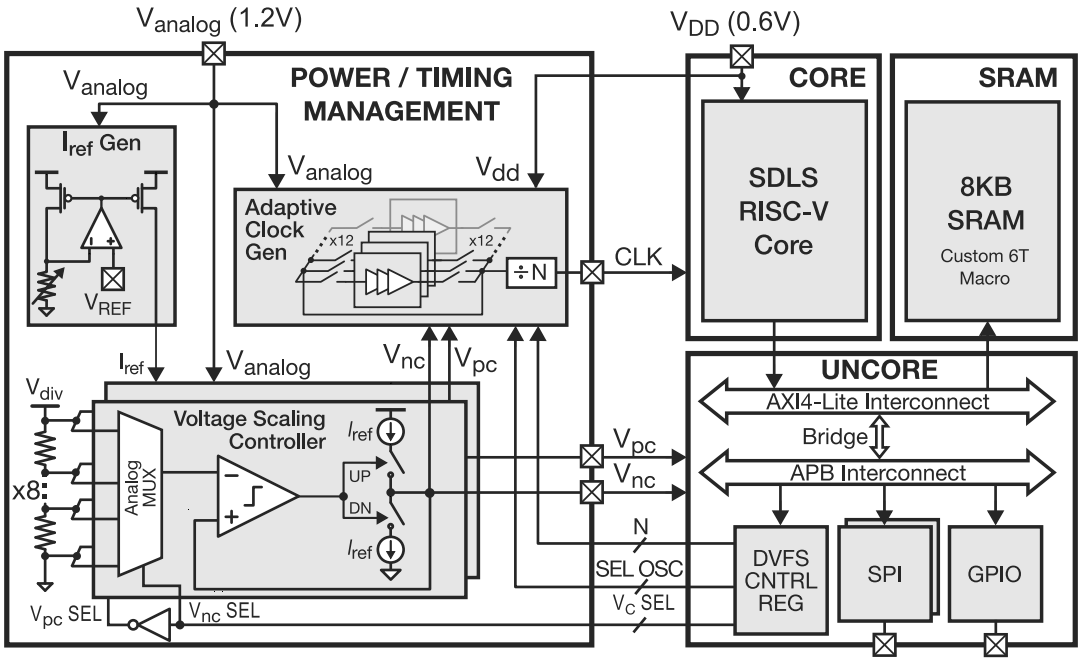

Fig. 2.34. Architecture of the SDLS-based RISC-V microprocessor. The core domain includes the RISC-V core, while the uncore domain includes interconnects, registers, and peripherals. A power and timing management subsystem includes a programmable reference current generator, the adaptive clock generator, and the voltage scaling controller

contains a positive-edge D Flip Flop with asynchronous reset, a logic buffer (BUF), inverter (INV), positive-edge latch (LAT), 2-input MUX, 2-input NAND, 3-input NAND, 2-input NOR, 3-input NOR, and an AOI22. This selection of cells (particularly the AOI22) was chosen by first synthesizing the RISC-V with a commercial standard cell library and evaluating which cells were most commonly used. It would not be feasible for this study to design a fully comprehensive library, but with this selection of gates, the original netlist was duplicated with 83% overlap. That is, only 17% of the gates in the original optimized netlist do not have logical equivalents in the SDLS library. These remaining gates can simply be constructed using a combination of existing gates (the NAND2 is a universal gate that can be used to make any other gate). The core, uncore, and SRAM all use a 0.6V nominal supply voltage. The power and timing management sub-system, consisting of the tunable reference current generator, voltage scaling controller, and an adaptive clock generator, operates at a 1.2V supply.

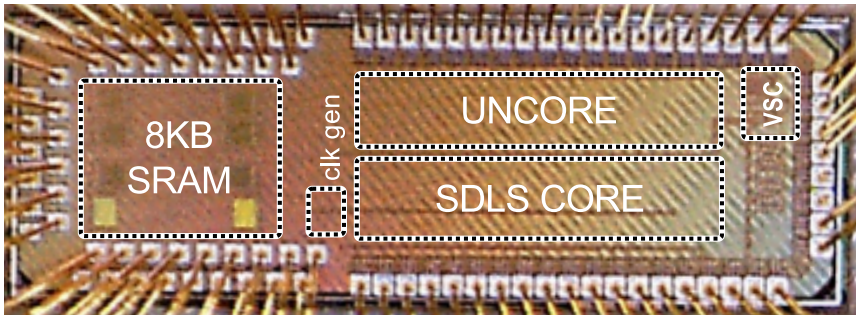

Fig. 2.35. Annotated micrograph of the SDLS microprocessor fabricated in 65-nm CMOS.

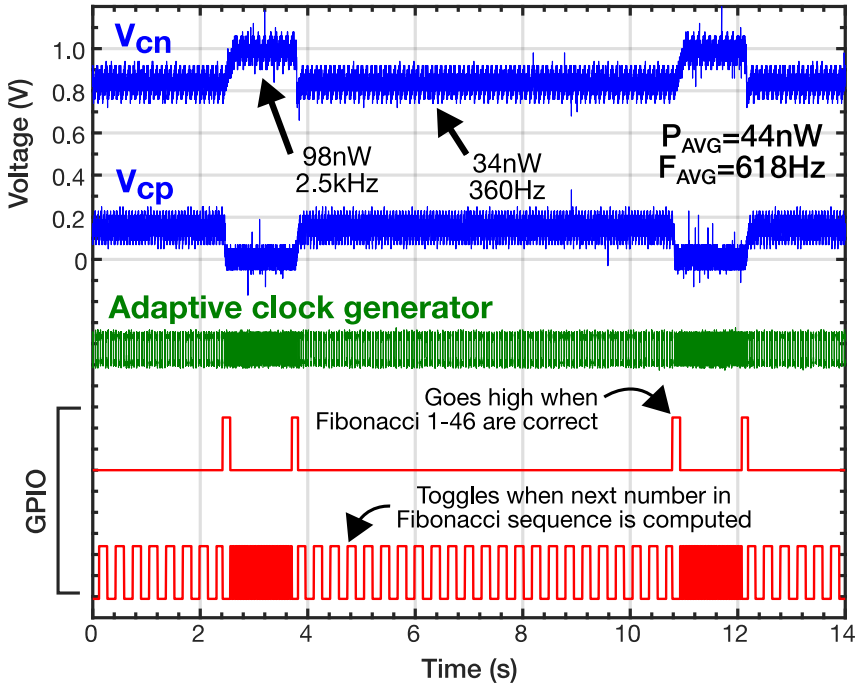

Fig. 2.36. Oscilloscope measurement of runtime DVFS with core dithering between two V_c *SEL* modes while computing the Fibonacci sequence. GPIO output bits indicate when a new Fibonacci number is computed and when the total sequence (numbers 1-46) are correct.

2.3.4 Measurements

The SDLS microprocessor test chip was fabricated in a 65nm low-power process. Fig. 2.35 shows an annotated chip micrograph. The core consumes an area of 0.045mm^2, while the uncore consumes 0.37mm^2 and the VSC and ACG consume 0.05mm^2 and 0.002mm^2, respectively. Fig. 2.36 shows oscilloscope measurements of V_{nc}, V_{pc}, the adaptive clock output, and GPIO bits during runtime DVFS while executing a self-checking Fibonacci sequence program. The program was written to compute the first 46 numbers in the Fibonacci

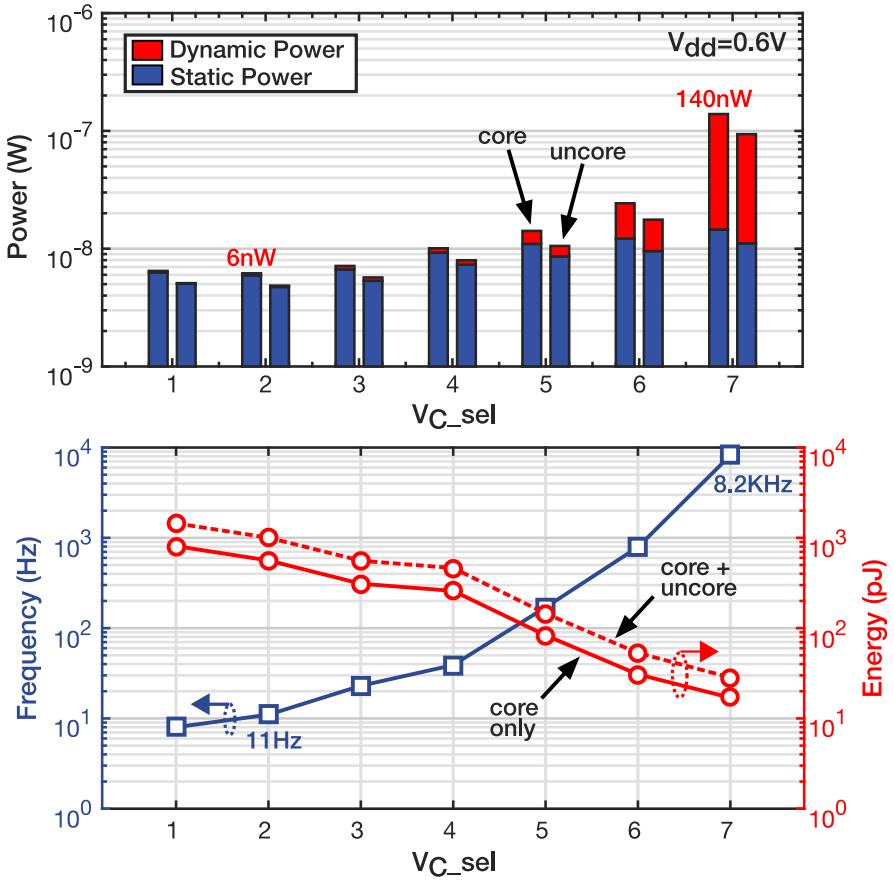

Fig. 2.37. Performance of the SDLS core and uncore at 0.6V V_{dd} across all possible DVFS modes, obtained by sweeping V_c *SEL* across all possible values.

sequence and toggle a GPIO bit after each successive number is solved. Once the 46th Fibonacci number is reached (1836311903_{10}), it is checked against a predefined value initialized in the program memory and if the result is correct then a separate GPIO bit will be briefly pulsed high and the core will transition to a different performance mode (adjust V_c *SEL* and the ACG signals) and repeat the above procedure. The entire program, including DVFS instructions, was written in C and compiled with a standard open-source toolchain. The measurements from this test show that the core executes the program successfully while dynamically shifting between two modes that achieve 98nW power at 2.5kHz clock frequency and 34nW at 360Hz clock frequency, respectively. To fully evaluate the SDLS core across all supply voltages and DVFS modes, a simpler GPIO bit toggling program is used. Fig. 2.37 shows the measured power, energy, and frequency of the core and uncore for each of the possible DVFS modes, obtained by sweeping V_c *SEL* from its minimum to its maximum

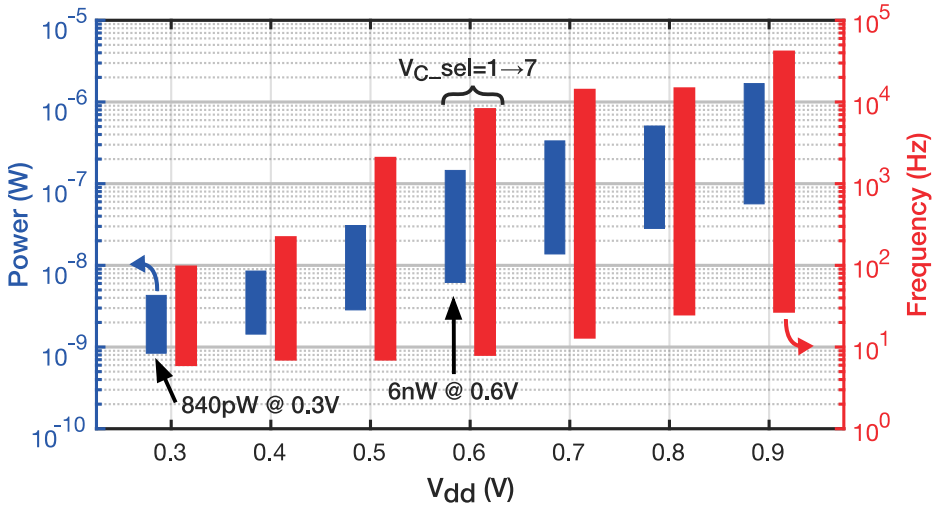

Fig. 2.38. Performance scaling ranges (core only) achievable through DVFS (sweeping V_c *SEL*) at each supply voltage point.

value, while running at the nominal V_{dd} of 0.6V. At this V_{dd}, the SDLS core consumes 6nW total power to run at 11Hz in the minimum DVFS mode and 140nW total power at 8.2kHz in the maximum DVFS mode, while the uncore consumes slightly less power increasing from 4.86nW to 93.9nW total power. The core and uncore are both dominated by static power at the slower DVFS modes, with dynamic power ramping up significantly the the upper 2-3 modes. The speed increase of the core outweighs the increase in total power as the DVFS mode is increased, so the total energy per clock cycle of the core decreases. Fig. 2.38 shows the achievable core power and frequency range from DVFS at each V_{dd}. That is, for every V_{dd} point, V_c *SEL* is swept from its minimum to maximum value, and the speed and power limits of the SDLS core are recorded. The minimum achievable core power is 840pW, which occurs in the minimum DVFS mode at 0.3V V_{dd} while running at 6Hz. At this point, the uncore domain consumes 690pW and the clock generator consumes 130pW, bringing the total minimum system power to 4.66nW after the addition of the 3nW VSC. While most processors require high operating frequency (and therefore high power) to reduce energy consumption, the SDLS core achieves a competitive minimum energy of 13.4pJ/cycle at 0.5V V_{DD} while running at 2.07kHz in the maximum DVFS mode, consuming just 27.9nW. Increasing V_{dd} allows the core to reach higher frequencies (up to 41.5kHz when V_{dd}=0.9V and V_c=1.1V) during DVFS, but results in high dynamic power due to the SDLS gates having a higher intrinsic

	This Work	[2]	[1]	[60]	[3]
Technology	65nm	180nm	180nm	90nm	130nm
Architecture	RISC-V	MSP430	ARM Cortex M0+	ARM Cortex M3	8-bit
Logic Family	SDLS	Dual-Mode DLS	DLS	static-CMOS	static-CMOS
Area (mm^2)	0.87	5.33	2.04	7.18	0.93
Operating Voltage (V)	0.3–0.9	0.2–1.1	0.16–1.15	0.5–1.0	0.45–0.9
Performance Scaling	Modified DVFS	Dual-Mode	None	DVFS + AVS	None
Scaling Granularity	7-point + dithering	2-point	None	40-point	None
Operating Frequency	6Hz–41.5kHz	2Hz–2.8MHz	2Hz–15Hz	1MHz–20MHz	40kHz–3MHz
Minimum Active Power	4.66nW @ 0.3V, 6Hz	595pW[1]@ 0.45V, 2Hz	127.1pW[3]@ 0.55V, 2Hz	34nW[1]@ 1.0V, 0Hz[2]	100nW[1]@ 0.45V, 40kHz
Minimum Energy	38.8pJ @ 0.5V, 2.07kHz	14pJ[1]@ 0.45V, 19kHz	44.7pJ @ 0.55V, 7Hz	23pJ @ 3V, 5MHz	2.8pJ @ 0.35V, 105kHz

[1] includes memory
[2] retention mode w/o power gating
[3] extracted from power breakdown

Table 7. Performance summary of SDLS RISC-V system design (includes core, uncore, VSC, and ACG) with comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (09/18/2019), and more recent works are not shown here. Reference publication for this work is [DT17].

gate capacitance than static CMOS gates. Table 7 summarizes the performance of the SDLS RISC-V core and compares it with other state-of-the-art microprocessors implemented with other logic styles.

2.3.5 Conclusion

This section discussed a custom Scalable Dynamic Leakage Suppression (SDLS) standard cell logic family designed to enable constant-V_{dd} performance scaling at ultra low-power levels. The SDLS logic family was implemented as RISC-V microprocessor in 65-nm CMOS and co-designed with a Voltage Scaling Controller (VSC) and Adaptive Clock Generator (ACG) to enable fully integrated "modified-DVFS" at any V_{dd}. With a 0.6V V_{dd}, the SDLS microprocessor can scale its performance from 6nW at 11 Hz to 140nW at 8.2 kHz. Overall, the core design achieves performance limits of 840-pW minimum power and 13.4-pJ minimum energy. This performance does a successful job at bridging the gap between traditional DLS logic and static-CMOS logic, depicted in Fig. 2.27.

3 Memory

3.1 Introduction

Memory is a necessary companion for digital processors to store instructions and data. IoT applications typically do not require large memory capacities beyond tens of kBs, so the entire system memory can be implemented on chip using a Static Random Access Memory (SRAM). SRAMs are volatile but dynamically stable in that they require power but not any additional attention (such as a refresh) to maintain their data for any long period of time. The prevailing method for storing a single bit of data within an SRAM array is with a 6-transistor (6T) bitcell, shown in Fig. 3.1. This design uses two cross-coupled inverters that drive each

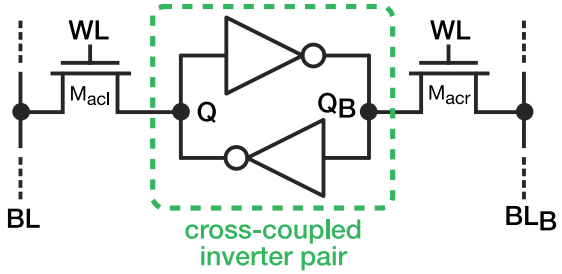

Fig. 3.1. Traditional 6T SRAM design (a), and implementation of 6T bitcell using DLS inverters (b).

other to two opposite voltages present at the internal nodes Q and Q_B. If Q=V_{dd}, then Q_B will equal 0, and the inverters will work together to maintain this state. In this state the bitcell is said to be storing a data '1'. Conversely, if Q=0 and Qb=V_{dd}, the bitcell is said to be holding a data '0', and the two inverters will similarly work to maintain this state. To "write" a data value into this bitcell, manual intervention is required to effectively force Q and Q_B to the desired values by overpowering the cross-coupled inverters. Once Q and Q_B are forced to the desired values, the cross-coupled inverters will then lock into this state and begin holding the data. This process is accomplished by charging the bit lines (BL and BL_B) to the desired (complementary) voltages, and then enabling the two access transistors (M_{acl} and M_{acr}) by applying a high voltage to the Word Line (WL). Assuming the BL drive strength is designed properly in accordance with the internal strength of the cross-coupled inverters, the

voltages Q and Q_B will be written to the desired voltages. To read the data value out of the cross-coupled inverter pair, the BLs are pre-charged (not actively driven) to V_{dd}, and then the access transistors are enabled as in a write operation. Whichever internal voltage is equal to V_{dd} will leave its associated BL unaffected, while the internal node being kept at 0V will begin to discharge its associated bitline. The differential voltage of the BL pair is then sensed with an amplifier (sense amplifier) at the end of the BL, and the polarity of the voltage reveals whether the bitcell is holding a 0 or a 1. Note that this operation should be non-destructive in that it doesn't corrupt the states of Q and Q_B during the operation. Just as digital logic gates (e.g., the inverter) are assessed for sufficient NMs as in Section 2.2 to guarantee reliable operation, the SRAM bitcell NM, also known as the static noise margin (SNM), needs to be carefully designed for successful and reliable operation [61]. Unlike a standalone digital inverter, the SRAM bitcell NM convolves not only the switching characteristics of both inverters but also the operating mode (hold, read, or write) of the overall bitcell. The hold state generally leads to the most simple NM characterization, while an active read operation deteriorates the NM due the (skewed) parasitic effect of the pre-charged BL on the VTC of one of the inverters. Likewise, during a write operation, a special case of the NM is defined such that a negative NM is required in order to guarantee the ability for the BLs to overpower the inverters and force Q and Qb to the desired voltages. There is a very large body of research focusing on the modeling and statistics of these NMs to improve the reliability of SRAM operation.

The discussed 6T bitcell design is tiled by the thousands to create many kB of data storage. While the overall resulting design, or "macro" contains peripheral circuits such as the sense amplifiers and address decoders to control the BLs and WLs, the most significant contributor to the total power is the collection of cross-coupled inverters. For a 6T SRAM macro with N total bits of storage, there are 2N inverters. Considering that only a very small number of these inverters are being switched at any given time, the dynamic power is often negligible compared to the leakage power, especially in IoT nodes that do not operate at high speeds. Therefore, the total power of an SRAM in standby or low-frequency operation can be approximated by

$$P_{total} \approx I_{leak} 2 N V_{dd} \tag{35}$$

3.1.1 SRAM power reduction

Significant efforts have been dedicated towards reducing standby power of SRAM by focusing on new operating schemes as well as circuit-level details. A popular operational scheme for reducing power is to reduce the SRAM supply voltage when it is not being used, resulting in a "sleep" or "shutdown" state [62, 63]. This is similar to DVFS in a digital circuit, where decreasing V_{dd} decreases both the power (due to Ohm's law) and the leakage currents of the transistors (due to DIBL). The amount of supply voltage reduction depends on whether the SRAM needs to retain its data during standby. If data retention is not necessary, the supply voltage can be completely cut off. However, the ability to retain data is usually desired, so the minimum V_{dd} of an SRAM must still guarantee adequate NMs [64]. If the maximum performance requirements (read and write speeds) of the SRAM are not very high, it can be designed to permanently operate at a low V_{dd} rather than only scaling down to a low V_{dd} during standby [65]. The low-V_{dd} Schmitt trigger logic (discussed in Section 2.1.2 and published in [44]) has also been used to implement an SRAM bitcell that can operate at very low supply voltages [66]. Additional efforts have been made to facilitate as much V_{dd} reduction from the bitcell as possible by modifying the peripheral circuitry to suppress leakage currents on the BLs that skew the NMs [67] or improving the reliability of the sense amplifiers when the differential BL read voltages become very small [68].

Regardless of any combination of operational or architectural techniques that are used, the leakage floor of existing SRAM arrays is ultimately still limited by the leakage of the cross-coupled inverter pairs. The most successful attempts to reduce the leakage of the cross-coupled pairs have used traditional digital circuit leakage reduction techniques (listed in Section 2.1.2) such as using High Threshold Voltage (HVT) transistors, long channel length devices, Reverse Body Bias (RBB), and stacking transistors [4, 5]. These techniques are quite effective in older technologies in which they are demonstrated, but can increase bitcell area and complexity and can have decreasing effectiveness in newer technologies due to the increasing relevance of gate leakage and the decreasing strength of the body effect.

3.2 DLS SRAM Design

This section describes the design and implementation of an SRAM bitcell based on the DLS logic described in Section 2.2. As the DLS inverter is functionally equivalent to the traditional static-CMOS inverters that are typically SRAM bitcells, a new "6T" bitcell can be designed that uses DLS inverters to form the cross-coupled inverter pair. The reduced leakage current of these DLS inverters will have a direct impact on the power limit shown by (35). Just as in traditional 6T SRAM bitcells, the new DLS SRAM bitcell should achieve the necessary NMs to guarantee reliable read, write, and hold operations. The reduced drive strength of DLS inverters (see Section 2.2.2) means that the cross-coupled pair is much more susceptible to NM degradation, especially during reads, than a traditional static-CMOS 6T bitcell. This section will review several design techniques to enable successful operation of the DLS bitcell, and the resulting design is implemented in 65-nm CMOS and verified with experimental results. The reference publications for the material in this section are [DT1, DT7]

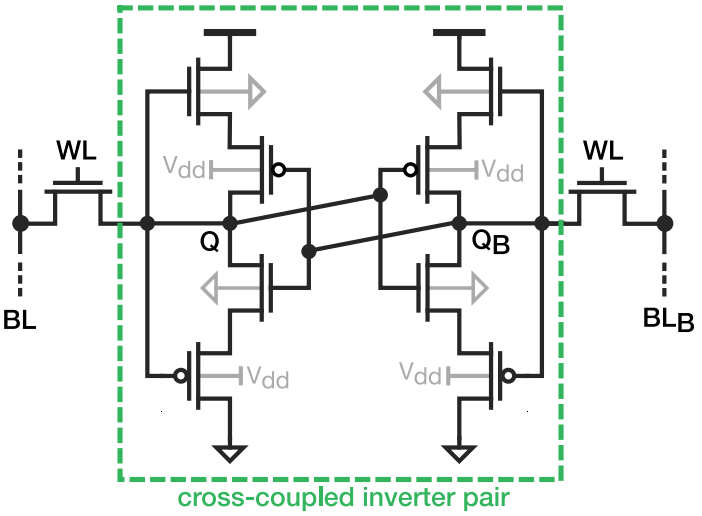

Fig. 3.2. Concept of 6T SRAM bitcell using DLS inverters (b).

3.2.1 DLS Bitcell

Fig. 3.2 demonstrates the concept of using DLS logic in a 6T bitcell. Functionally, this is identical to a classic 6T cell where the two cross-coupled inverters create the data storage voltages Q and Q_B. All transistors in the DLS inverters are implemented with thick-oxide

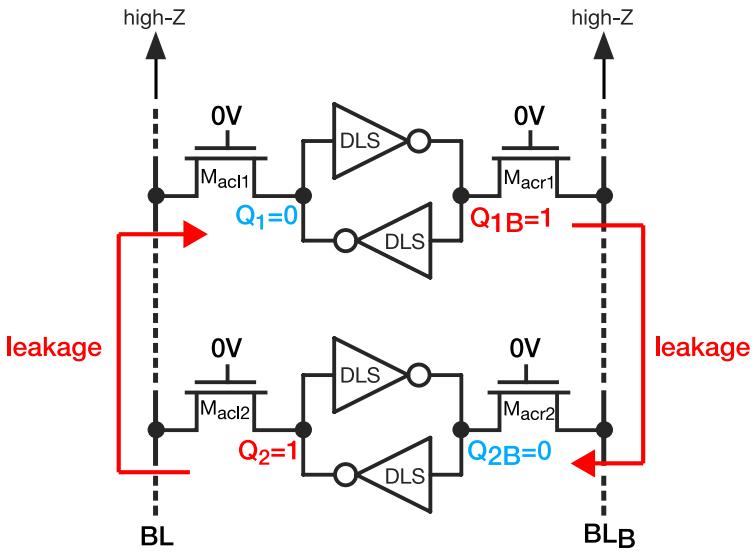

Fig. 3.3. Potential for leakage currents through the BL to disrupt the stored data in adjacent DLS bitcells.

I/O devices to ensure that gate leakage remains negligible relative to subthreshold leakage. The first consideration for successful operation is the hold stability, when access transistors are turned off. In this state, the Q and Q_B nodes are generally safe since the cross-coupled inverters can overpower the leakage through the access transistors and maintain the data state if the BL is floating with high impedance looking towards the drivers. However, when multiple DLS bitcells are combined onto a BL, as shown by Fig. 3.3, adjacent Q and Q_B nodes can potentially be disrupted by charge that leaks out onto the BL from one bitcell and into an adjacent bitcell. In the example shown, this occurs when two adjacent bitcells are holding a data 0 (Q_1=0 and Q_2=1), where the only thing that stands in the way between them is the series leakage of M_{acl1} and M_{acl2}. Generally, to guarantee hold stability, we can say that the internal drive strength of the DLS inverters should be able to overpower the leakage of the series access transistors, which can be modeled exactly like the stack effect in Section 2.1.2. A more "worst-case" scenario is when one bitcell on the BL is being written to, so the BLs are actively driven to either 0V or V_{dd}. For the bitcells on the BL that are not being accessed (their WL is at 0V), the leakage through their access transistors can be represented by their subthreshold current when V_{ds}=V_{dd} and a V_{gs} of 0V. We can then adopt

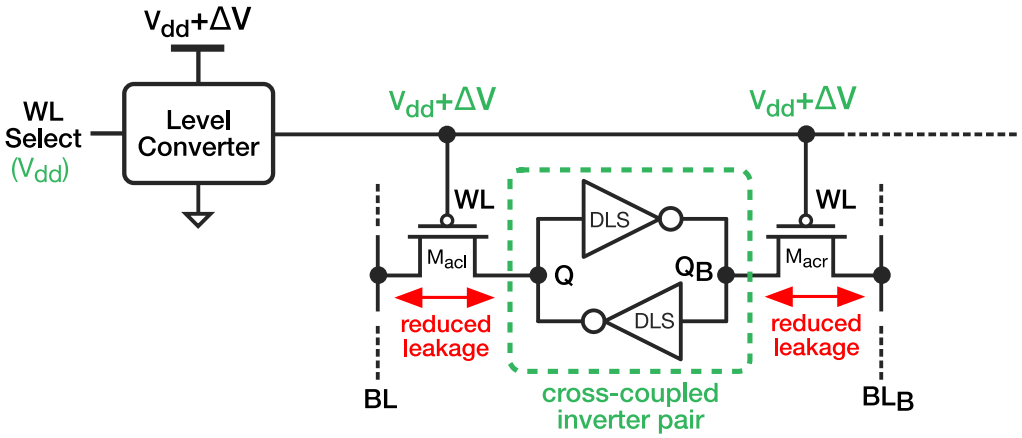

Fig. 3.4. Implementation WL overdrive to suppress leakage current through the access transistors.

the requirement from (26) to assert that

$$I_{ds,Mnh}\big|_{V_{gs}=0,V_{ds}=\sim kT/q} \gg I_{ds,Macl1}\big|_{V_{gs}=0,V_{ds}=V_{dd}} \tag{36}$$

where M_{nh} corresponds to the DLS PU transistor in the lower bitcell that is responsible for keeping Q_2 charged to V_{dd} (data 1). Based on the characteristics of M_{nh} and M_{acl}, this condition is not easily met without further compensation to suppress the leakage of the access transistors. One approach to reducing the access transistor leakage is to use a negative V_{gs} to place the transistors in a super-cutoff region which is the same mechanism used intrinsically by the DLS logic style. Using the default NMOS access transistors, this would require a negative WL voltage with respect to ground [63], which can complicate the design. To address the need for reduced access transistor leakage, M_{acl} and M_{acr} are instead implemented with PMOS transistors, as shown in Fig. 3.4. Now, rather than needing a voltage less than 0V to underdrive the gate, a positive voltage above V_{dd} can be used to overdrive the gate ($V_{sg} < 0$) to suppress the leakage just as in M_{pf} from in a DLS inverter design. Similar to the transistors in the cross-coupled DLS inverters, these access transistors are implemented with thick-oxide I/O devices to ensure that gate leakage is insignificant, which is even more relevant here since the gate voltage will be boosted higher than V_{dd}. Given a nominal supply voltage V_{dd}, the WL voltage can be boosted by a small amount ΔV to achieve this result. In this work, a ΔV of 0.3V is sufficient to reduce the leakage of the access transistors to a level below the requirement in (36). Fig. 3.5 shows a Monte Carlo simulation result of the

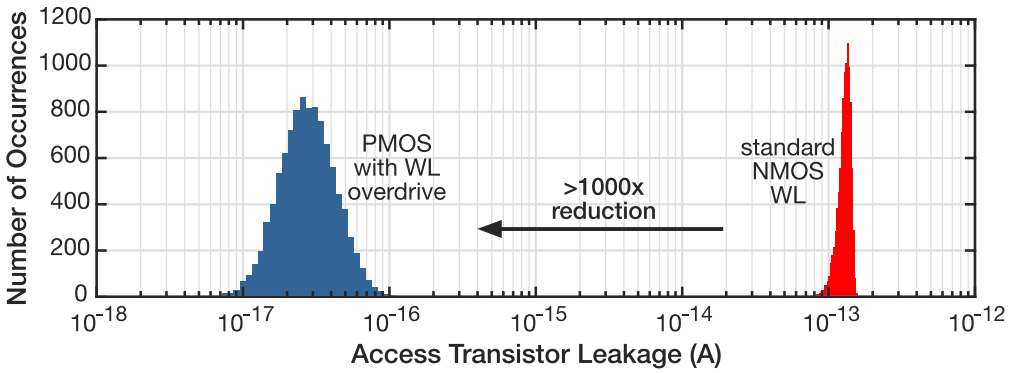

Fig. 3.5. Monte Carlo simulation of the leakage reduction through the access transistors for the traditional NMOS access transistors and the utilized PMOS access transistors with overdrive WL voltage.

leakage currents of the access transistors, comparing the traditional NMOS-based access transistors and the proposed PMOS-based access transistors with overdriven WL voltage. With the overdriven PMOS access transistors, leakage is reduced by a factor greater than 1000. Assuming the necessary $V_{dd} + \Delta V$ voltage can be obtained reasonably easily from within an IoT SoC (where the exact value of ΔV can be somewhat flexible based on what voltages are available), the only additional modification required here is a level converting circuit to convert the WL signal (now called WL Select) from the V_{dd} domain to the $V_{dd} + \Delta V$ domain. A key requirement for this level converter circuit is to be both compact and very low power so as to not mitigate the power savings of the DLS bitcell or increase the required chip area to a prohibitively large value. To meet these needs, a DLS-based level converter is proposed, which will be discussed next in Section 3.2.2. The WL overdrive technique satisfies stability requirements of the bitcell for any case when the access transistors are meant to be turned off, meaning that the data will be properly retained when the array is in standby or when adjacent cells are being access for reads or writes. The second case to be considered for successful operation is stability when the access transistors are meant to be turned on, which only occurs when the bitcells is being read from or written to. In the case of a write operation, which is inherently data-destructive, it can simply be ensured that the BL drivers and access transistors enable enough current to pour in and out the bitcell (at nodes Q and Q_B) to overwrite the data state. During a read operation, the access transistors are similarly turned on, but now the bitcell is meant to sink current from the BL through the cross-coupled inverters. If the current flowing into Q (or Q_B, depending on which node is

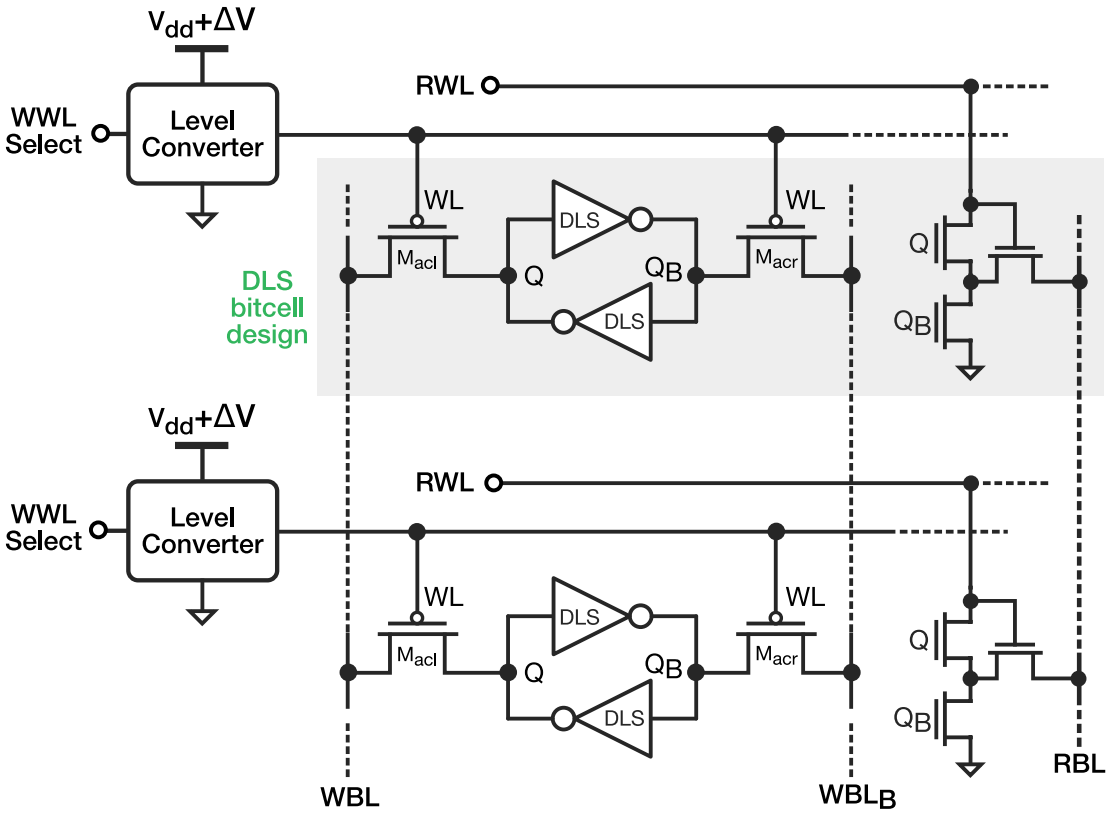

Fig. 3.6. Implementation of DLS SRAM array using DLS level shifter for WL boosting.

holding a 0) from the pre-charged BL is higher than the current being discharged from Q/Q_B via the the DLS inverters, then the voltage on Q/Q_B can rise enough to flip the state of the data inside the bitcell, resulting in data corruption. The solution to enable non-destructive reads is the addition of a read port, as shown in Fig. 3.6. This single-ended port uses three extra transistors to electrically buffer the data state out to an independent Read Bit Line (RBL) whose value will reflect the internal Q node. A new Read Word Line (RWL) is used to enable the read port, while the original WL now becomes a Write Word Line (WWL) for write-only operations. Likewise, the original BLs are now Write Bit Lines (WBLs).

With the addition of the read port, the completed DLS bitcell is now fully operational, capable of reading, writing, and holding. As mentioned in Section 3.1.1, low-V_{dd} operation is critical for SRAM power reduction. This is completely aside from the reduced leakage currents of the DLS inverters, since lower V_{dd} will continue to reduce power due to Ohm's law. One distinct advantage of using DLS logic is the fact that its NMs are greater than $V_{dd}/2$, giving it greater hold stability than a static-CMOS inverter. Therefore, the DLS SRAM bitcell can be

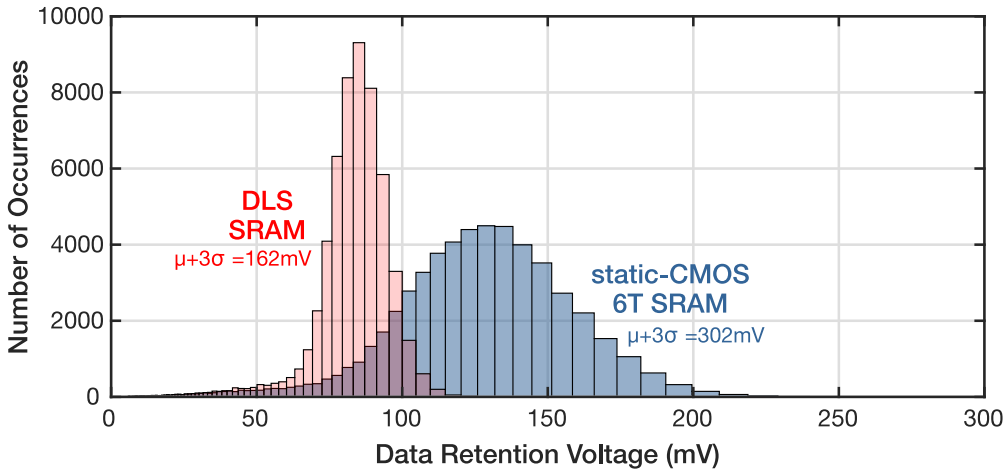

Fig. 3.7. Monte Carlo simulation of the Data Retention Voltage (DRV) of a traditional 6T static-CMOS bitcell and the proposed DLS bitcell, showing an improvement in minimum retention V_{dd} from the DLS bitcell.

dropped to extra low supply voltages while still being able to retain its data due the improved noise immunity from its large NMs (NM deterioration at low V_{dd} is what causes inverters to eventually fail to operate). The below-average SMs of DLS gates do not matter here since it is not desired to be able to switch the gate in a retention-only state. This minimum supply voltage at which the bitcell can retain its data (but not necessarily operate at) is called the Data Retention Voltage (DRV). Since the DRV only guarantees that data will be retained, the V_{dd} must be first raised back to the nominal value before performing a read or write. Fig. 3.7 shows the simulated Monte Carlo values of the DRV for the traditional 6T static-CMOS bitcell and the final DLS bitcell with word line overdrive and read port. As expected, the DLS bitcell can retain its data at a noticeably lower supply voltage, and additionally sees less variation across the Monte Carlo sweep. The conservative approach to assigning a DRV value to an SRAM macro is to take the mean (μ) and add the 3-σ variation to it to define the lowest V_{dd} that will guarantee that all bitcells retain their data. For the DLS SRAM, the final predicted DRV value is 162mV, while the static-CMOS 6T bitcell gets a DRV of 302mV.

3.2.2 DLS level converter

To overdrive the WL signal for the PMOS access transistors, a level converter can be used to shift the voltage of the incoming WL select signal (V_{dd} domain) to a voltage slightly higher than the rest of the DLS array ($V_{dd} + \Delta V$). This would require one level converter per word

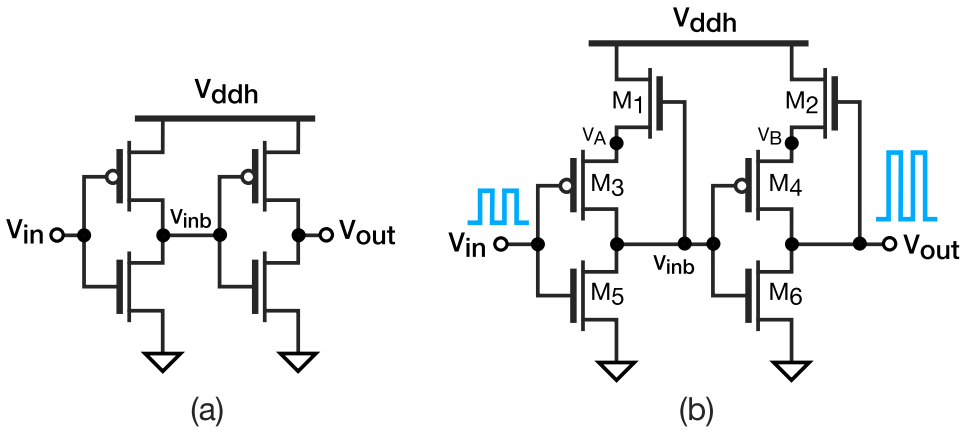

Fig. 3.8. Schematic of a traditional static-CMOS digital buffer (a) and schematic of DLS-based level converter (b).

line, and when considering the state-of-the-art for level converter designs [69–71], this would add significant leakage and area to the SRAM. To solve this problem, a DLS-based level converter is proposed that achieves low area and significantly less leakage than existing level converters. Fig. 3.8(b) shows the schematic of the proposed level converter, which follows the architecture of a traditional digital buffer consisting of two cascaded inverters (Fig. 3.8(a)). The traditional digital buffer on any arbitrarily high supply voltage V_{ddh} (1.2V is used here as an example) may be modified to operate from a 0.3V (V_{ddl}) input signal by strengthening the NMOS and weakening the PMOS devices in each inverter such that their currents are equal at a input logic-high threshold voltage $V_{ih} \approx V_{ddl}/2$, but this leads to several hundred nA of static power dissipation due to short-circuit current through the PMOS when $V_{in} = V_{vddl}$ or high subthreshold leakage current through the strengthened NMOS when V_{in}=0V (Fig 3.9). To preserve the simplicity of the skewed-inverter approach but eliminate the sources of static power dissipation, a modified inverter design based on the DLS technique can be used that alters the Pull-Up (PU) network by adding an NMOS header transistor between the PMOS transistor and the V_{ddh} supply rail. The gate of the NMOS header is controlled using feedback from the output of the inverter such that the header is enabled (shorted) when the pull-up network is active, and cutoff with a negative V_{gs} (ultra low leakage) when the PU network is inactive. As a result of the DLS behavior, it is possible to skew the PU and PD networks to achieve the desired V_{ih} without incurring static power penalties since the NMOS can be sized to have low leakage (low static power when V_{in}=0V) while still being

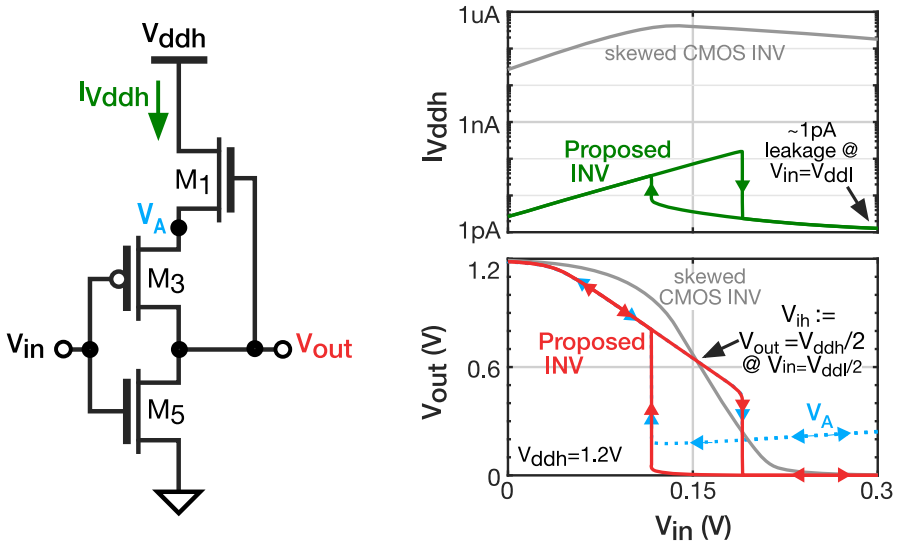

Fig. 3.9. Schematic of DLS-based inverter used in the level converter design and simulated DC transfer characteristics at V_{ddh}=1.2V with comparison to a traditional static-CMOS digital inverter with skewed PU and PD sizes.

strong enough overpower the DLS-based PU network when $V_{in} = V_{ddl}$, and when $V_{in} = V_{ddl}$ the DLS PU network will suppress short-circuit current that would normally be flowing through a traditional static-CMOS PU network. Fig. 3.9 shows the schematic and simulated DC transfer characteristics of the proposed DLS-based inverter, which has a characteristic that is nearly identical to the standard DLS inverter VTC in Fig. 2.7 with two key exceptions that are both caused by the removal of the PMOS footer. First, the absence of any feedback in the PD network prevents V_{out} from being held at V_{dd} as V_{in} increases, so V_{out} instantly begins to droop as V_{in} rises. Second, the relatively increased strength of the PD network means that it does not take a very high V_{in} value to cause M_5 to overpower the PU network, which was the general approach for finding V_{ih} in (21). Another way to look at this, using (22), is that by deleting M_{pf}, we can treat its V_{th} as 0V, leading to a lower equivalent V_{ih}. This is the key mechanism that allows the design to function as a level converter via low input switching threshold voltages. To ensure that the inverter output transition occurs at a switching threshold V_{ih} that is compatible with the input domain V_{ddl}, we can follow the approach for the skewed inverter by choosing the relative sizes of M_1 and M_5 such that they are in equal contention when $V_{in} = V_{ih} = V_{ddl}/2$. Therefore, we equate the subthreshold currents of M_1 and M_5 assuming that $V_{in} = V_{ddl}/2$ and $V_A = V_{out} = V_{ddh}/2$. Using the standard

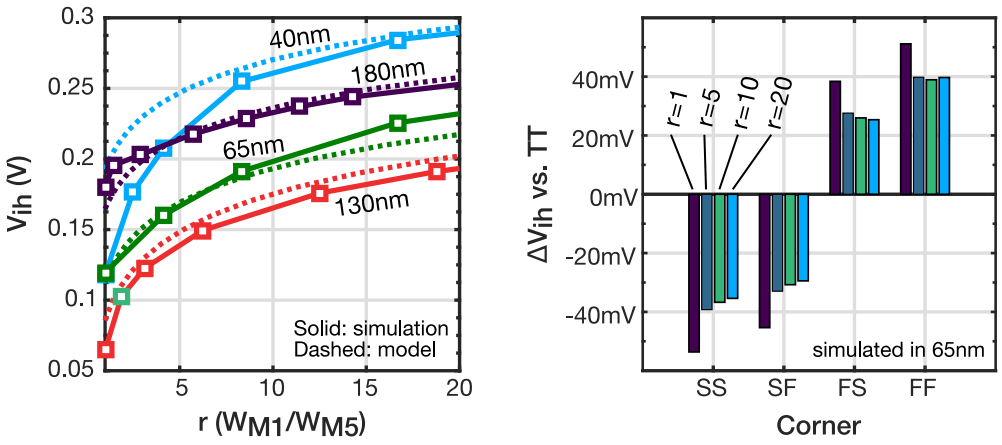

Fig. 3.10. Modeled and simulated input logic-high threshold voltage V_{ih} for the DLS level converter versus the sizing ratio its transistors, and the susceptibility of V_{ih} to process corner versus the nominal (TT) case.

subthreshold current equation from (14) and (for simplicity) assuming that any difference in the subthreshold swing factor n and DIBL coefficient η of M_1 and M_5 will be negligible. Then, we obtain a sizing ratio r

$$r = \frac{W_{M1}}{W_{M5}} = e^{\frac{(V_{th,M1} - V_{th,M5} + V_{ih})}{nV_T}} \tag{37}$$

If we assume M_1 and M_5 to both have the same threshold voltage, then a V_{ddl} of 0.3V ($V_{ih} = 0.15V$) requires M_1 to be sized around 73 times larger than M_5, which would occupy a significant area. Instead, M_5 is kept as a minimum-sized HVT device for low leakage, and M_1 is implemented with a LVT device. This is equivalent to the LVT-HVT DLS logic family shown in Table 4. The resulting r values are shown in Fig. 3.10 for several commercial technologies, and range around 1 to 5 for the targeted $V_{ih} = 0.15V$. Note that these results are independent of V_{ddl} and V_{ddh} as shown by (37). Lower r-values result in lower overall leakage due to smaller device sizes, but also slightly increases the susceptibility of V_{ih} to systematic process variation, assuming devices with different thresholds are equally subjected to the same process variations. Fig. 3.10 also shows the deviation of V_{ih} from the TT corner for different r-values ranging from 1 to 20. Since V_{ih} depends on M_1 and M_5 but not M_3, it is mainly sensitive to the NMOS corner. For r-values above 5, the variation in V_{ih} is limited to $\pm$40mV across all corners. Further descriptions and additional measurement data on this DLS-based level converter design can be found in [DT1].

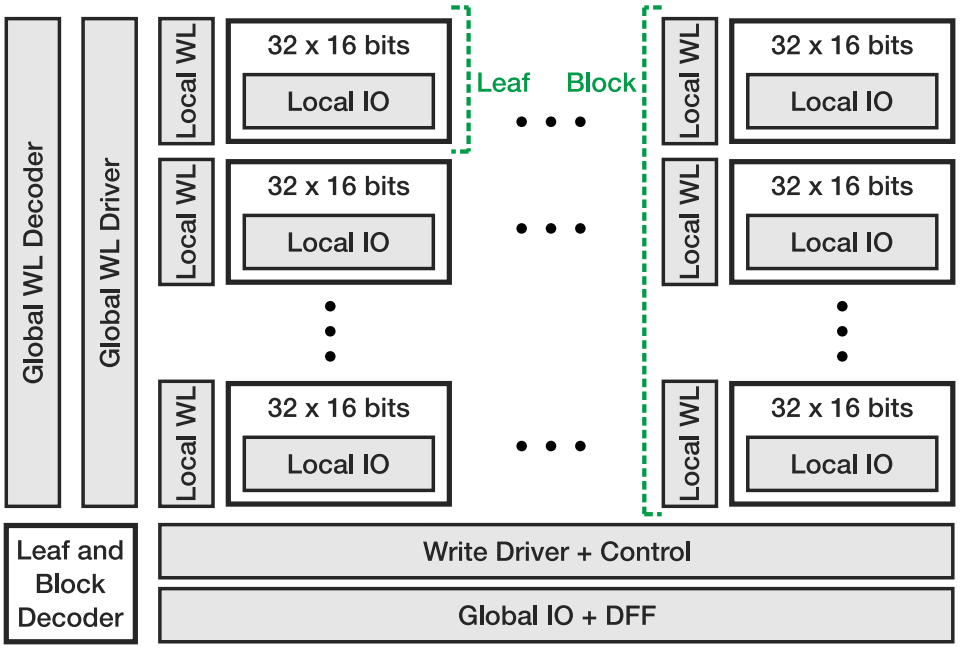

Fig. 3.11. Architecture of the 16kb DLS SRAM macro.

3.2.3 Measurement Results

The DLS bitcell design was implemented in a 16 kb macro based on the architecture shown in Fig. 3.11. The array is divided into leafs of 512 bits with local I/O and WLs. All of the peripheral logic is implemented with static-CMOS to minimize area and delay overhead, so separate power supplies were used for the bitcells and the peripheral circuits to measure their power consumption separately. More information on the architectural implementation can be found in [DT7]. Fig. 3.12 shows the die photo of the full SRAM macro, which consumes an area of 0.312mm^2. The DLS bitcell itself consumes an area of 12.44μm^2, leading to an array area efficiency of 67.8%. 10 dies were tested for their minimum DRV, showing a best-case DRV of 115mV and an average DRV of 156.6mV with a standard deviation of 26mV. Fig. 3.13 shows the measured leakage of the DLS SRAM. The leakage current per bitcell is measured by taking the leakage current of the entire array (excluding the periphery that is on a separate power supply) and dividing by the total number of bits. The use of thick-oxide I/O devices prevents leakage current from increasing at high V_{dd} as shown in Fig. 2.14. Instead, subthreshold leakage and junction leakage dominate the total power, leading to a characteristic that only slightly increases across V_{dd}. Compared with other

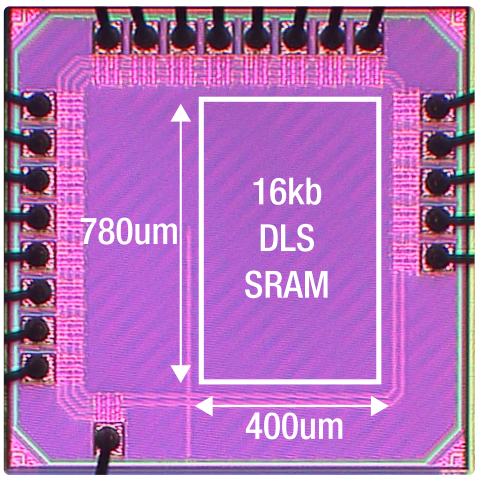

Fig. 3.12. Fabricated 16kb DLS SRAM test chip in 65-nm CMOS.

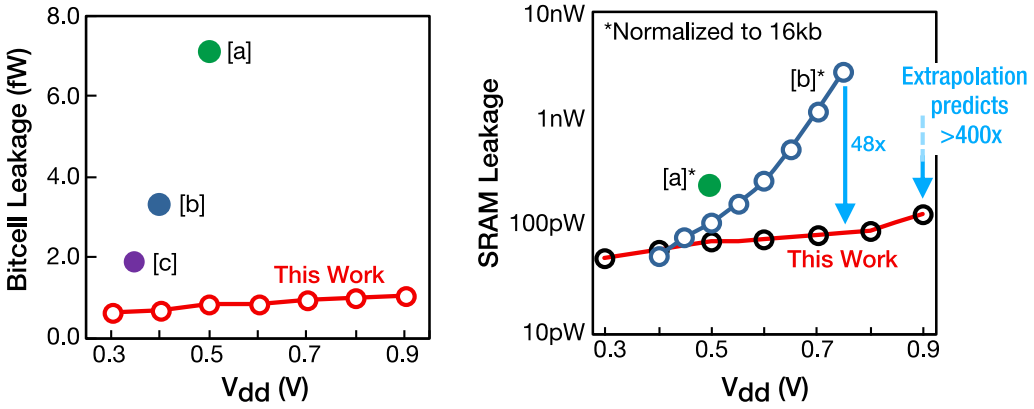

Fig. 3.13. Measured leakage current of the DLS bitcell and leakage power of the full SRAM macro across V_{dd}. Reference [a] corresponds to [3], [b] corresponds to [4], and [c] corresponds to [5].

low-power-oriented bitcell designs from recent literature, the DLS bitcell achieves the lowest leakage power at less than 1fW. Factoring in the power from the rest of the periphery, Fig. 3.13 also shows the total leakage for the full macro. Because the peripheral circuits all use traditional static-CMOS with relatively higher leakage, the full SRAM leakage power rises up to align more closely with the prior works whose peripheral circuits benefit from low subthreshold leakage due to the older technology nodes with larger devices and higher threshold voltages. Still, with the majority of the SRAM leakage being due to bitcell leakage, the total leakage power stays relatively constant across V_{dd}, improving on the prior work by nearly 50x at a 0.7V supply and by over 400x if the measurements are extrapolated to 0.9V, where the total leakage power of the DLS SRAM is 132pW. The benefit of this low

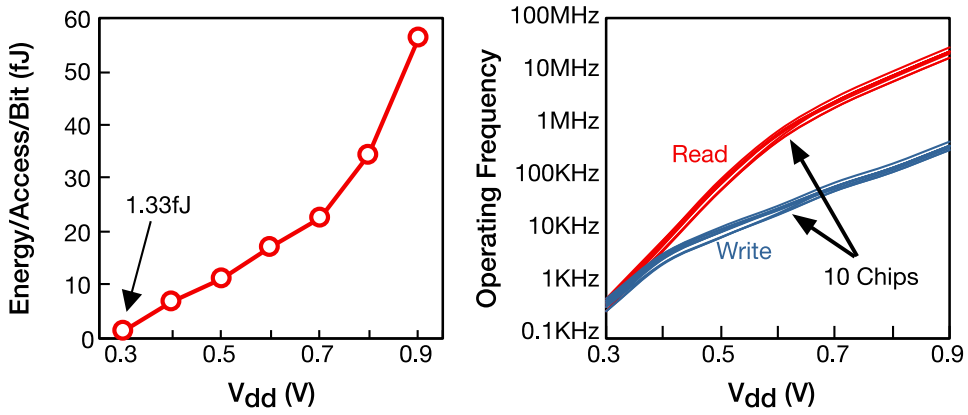

Fig. 3.14. Measured operating frequencies of the DLS SRAM (both reading and writing) and read energy (energy per access per bit) versus V_{dd}.

power operation at high V_{dd} is that the design does not need to run from a low-V_{dd} voltage regulator that adds inefficiency losses to the total power consumption. Instead, the design can run unregulated or at a higher regulated supply voltage. A key (somewhat unintended) benefit of using a read port with the DLS bitcell is that the weak internal drive strength of the DLS inverters does not limit the read access speed. Under a conventional BL sensing method, the DLS inverter would discharge the BL very slowly, adding a delay overhead to the the sense amplifier being able to detect the differential BL voltage. In combination with the traditional static-CMOS peripheral circuits that can operate much quicker than DLS logic, the read port allows data to be read fro the DLS SRAM relatively fast, reaching several kHz at low V_{dd} well over 1MHz at 0.9V. The effective dynamic energy, or energy per access, during the read operation is also shown, normalized per bit. At the minimum supply voltage of 0.3V, it takes only 1.3fJ per bit for a read access, increasing roughly quadratically with V_{dd} as expected based on (6). Write operations are relatively slower, still possible at several kHz at low V_{dd} and becoming limited at a few hundred kHz at 0.9V. Finally, the robustness of the DLS bitcell design is verified across a temperature range of 0°C to 60°C. Fig. 3.15 shows the measured access frequency versus supply voltage, which increases along with temperature. The dependence on temperature is greater at lower supply voltages than at high supply voltages. At the highest measured temperature (60°C), the minimum operational V_{dd} of the design increases by about 50mV or so. The total leakage current remains in the pW-range across the full V_{dd} range until the temperature reaches 40°C, at which point it

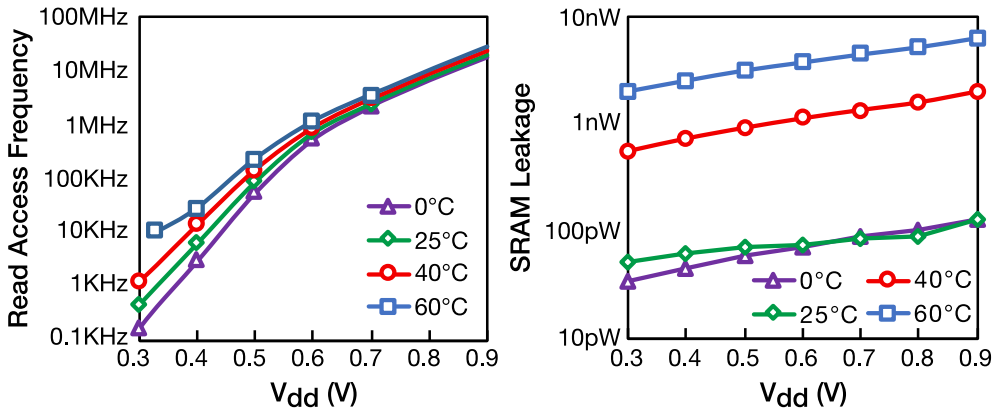

Fig. 3.15. Measured leakage power and read access frequency of the DLS SRAM versus V_{dd}, shown across a temperature range of 0°C to 60°C.

begins increasing to several nW. Table 8 summarizes the performance of the DLS SRAM.

	This Work	[3]	[4]	[5]
Technology (nm)	65	180	180	180
Bitcell Type	13T (DLS)	14T	10T	10T
Bitcell Area (μm^2)	12.44	40	17.48	17.48
Access Frequency	6.9kHz @ 0.4V	100kHz @ 0.5V	3.5kHz @ 0.35V	1kHz @ 0.3V
Write Frequency	3.6kHz @ 0.4V	Async	Multi-Cycle	Multi-Cycle
V_{min} (V)	0.3V	0.5V	0.35V	0.3V
Best DRV (mV)	115mV	–	–	–
Read Energy (fJ)	113	–	–	–
SRAM Leakage (pW)	51.87	22	80.53	–
Capacity	16kb	1kb	24kb	24kb
Area Array Efficiency	67.8%	64%	27%	27%

Table 8. Performance summary of DLS SRAM and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (6/16/2020), and more recent works are not shown here. Reference publication for this work is [DT7].

3.2.4 Conclusion

This section presented an SRAM bitcell design based on Dynamic Leakage Suppression (DLS) logic to enable low-leakage SRAM macro implementations. The general operating requirements of the DLS bitcell are reviewed, and several bitcell-oriented design modifications are proposed to enable the DLS bitcell to achieve successful operation. A 16kb SRAM macro is implemented with the DLS bitcell and fabricated in a test chip in 65-nm CMOS. Measurement results show that the DLS bitcell has the lowest leakage published to date amongst existing SRAM bitcells (to the best of current knowledge) with a value of less than 1fW/bit, and the full SRAM macro achieves less than 1nW of power from 0.3V to 0.9V.

4 Clocking Circuits

4.1 Introduction

Clock signals are necessary to drive the operation of many circuits and systems such as radios, wireline communication circuits, analog front ends, digital logic, and memory. The requirements for the clock signal can vary widely depending on the application requirements as well as the tolerance of the circuits that use the clock. For example, wireless communication standards can dictate the noise limits for clock signals used in radios. Clock signals used in timer circuits need to have good long-term stability, which dictates the temperature stability. In digital circuits, functional errors can occur if the clock signal has too much period jitter. Due to all these factors, the overall clocking requirements and prioritized metrics for an IoT system can vary widely, and it is not feasible to have a single universal clocking solution. However, most system clocking solutions are based on a few fundamental architectures that can be characterized by 1) whether they can be implemented fully on-chip, and 2) whether they are temperature compensated. Fig. 4.1 shows the most classic examples of these fundamental architectures: The RO which is on-chip but not intrinsically temperature compensated, the Relaxation Oscillator (RXO) and Frequency-Locked Loop (FLL) which are both on-chip and naturally temperature-compensated, and the Crystal Oscillator (XO) which is off-chip and temperature-compensated.

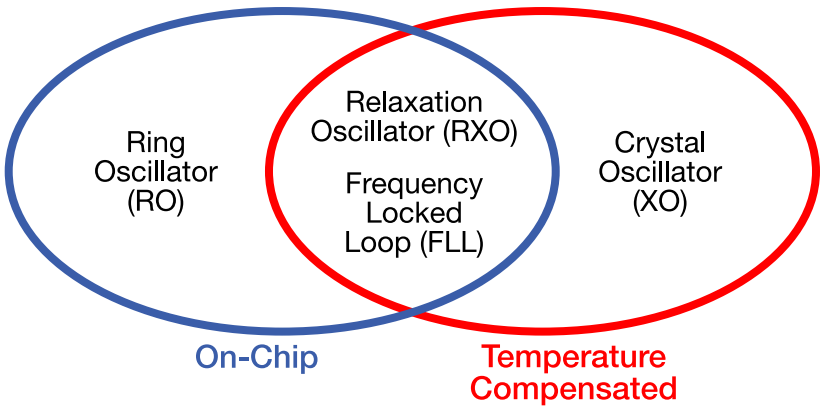

Fig. 4.1. Classification of common designs for clocking circuits.

In general, all clocking circuits operate by creating a time-varying output voltage whose frequency F is based on some physical reference unit of measure of time. If the physical reference is temperature-stable, then the output frequency of the clock should be similarly stable. Finding or creating a temperature-stable reference is the first challenge. On-chip, an RC time constant is often used since capacitors are usually very temperature-stable and resistors can be designed to cancel out first-order temperature dependencies, which will be shown later. An interface circuit such as an RXO or FLL is responsible for driving the RC delay to create an output clock signal. In the context of circuit design, this interface circuit is usually referred to as the actual "clock", whereas in signal processing and digital circuit operation, the "clock" instead refers to the generated frequency F. Using an on-chip reference helps reduce the area of the overall system and does not add any integration overheads, but the limited temperature stability of the on-chip passives can be an issue. Some techniques to further improve the resistor stability involve using advanced circuits for second-order compensation [72] or using an on-chip temperature sensor to dynamically tune the resistor value as temperature varies [DT18]. If off-chip components can be tolerated, a quartz crystal reference can be used with an XO to create a fixed frequency that has several orders of magnitude more temperature stability than on-chip oscillators. If no temperature compensation is needed, a much simpler on-chip design can be used such as an inverter-based RO whose physical reference is the propagation delay of a digital inverter. This propagation delay is intrinsically exponentially-dependent on temperature, but there have been efforts to compensate this as well [73]. A final consideration worth mentioning here is the frequency tunability of the different architectures. ROs have the greatest amount of frequency tuning, since the delay of an inverter is exponentially dependent on the supply voltage which is often used as the tuning knob. Temperature compensated oscillators like RXOs and FLLs can achieve some tunability by adjusting the RC values, but this can add some extra temperature variation. XOs can be selected from a broad frequency range, but are individually only capable of operating at one frequency. Outside of the core clock design, additional circuits such as counter-based frequency dividers or Phase-Locked Loops (PLLs) can be used to add tunability by multiplying or dividing the clock frequency, but these can also add extra power consumption and design complexity.

This dissertation focuses on on-chip oscillators since they are attractive solutions IoT nodes due to their low form factor and reasonable performance. For microcontrollers specifically, on-chip temperature-compensated oscillators can generally meet the temperature stability and noise performance needs, but reductions in power consumption are necessary to improve the ability for self-powered operation as will be shown next. As mentioned previously, two common architectures for on-chip clock generation are the RXO [DT18, 74–77] and FLL [78–84]. These architectures will be the focus on the remainder of this section.

4.1.1 The Relaxation Oscillator

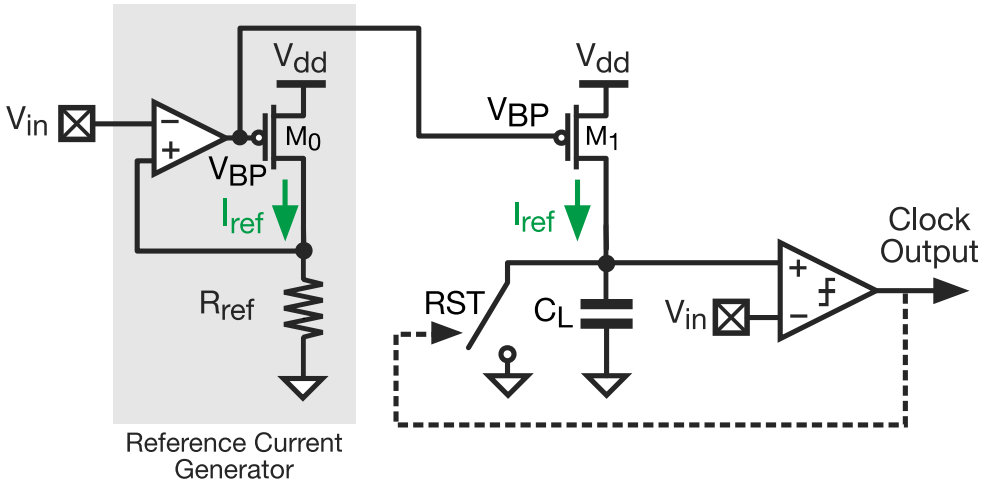

Fig. 4.2. Classic Relaxation Oscillator (RXO) architecture.

The RXO, shown in Fig. 4.2, has been a popular design for decades, and its modern design is captured well in [85]. It typically operates by transforming the reference resistor R_{ref} into a reference current I_{ref} which is mirrored from transistor M_0 to transistor M_1 and integrated as charge on a capacitor C_{L}. The reference current can be generated with an amplifier-less approach, such as a beta-multiplier circuit, or by using the approach depicted in Fig. 4.2 in which an amplifier is used to form a voltage-to-current (V-I) converter using an input reference voltage V_{in}. In the latter case, due to feedback from the amplifier that regulates its positive (+) input to V_{in}, the generated I_{ref} is simply equal to $V_{\text{in}}/R_{\text{ref}}$ as determined by Ohm's Law. A comparator monitors the rising voltage of the capacitor and toggles its output when the accumulated charge passes a threshold value, which in this case is the same V_{in} value.

Each comparator toggle resets the voltage on the capacitor and toggles the output clock signal of the oscillator. Applying nodal analysis to the C_L node provides the time required for C_L to charge to V_{in}, which then determines the period and frequency of the RXO:

$$F_{out} = \frac{1}{2 \times t_{charge}} = \frac{I_{ref}}{2 V_{in} C_L} = \frac{1}{2 R_{ref} C_L} \tag{38}$$

This simple equation for the output frequency shows that increasing either the resistor or capacitor value will provide a lower output frequency. Temperature dependence of the output frequency can now be simply expressed with (38) as long as the temperature dependencies of R_{ref} and C_L are known.

Power consumption in RXOs can be modeled as in other areas of this dissertation by considering both static and dynamic power contributions. Static power in the presented design is due to the I_{ref} that flows through various areas of the circuit. It is dissipated though both R_{REF} and C_L (a factor of $2^* I_{ref}$), and if the comparator is implemented as a continuous-time design, some multiple k of I_{ref} will also be used to bias the comparator. Then, the static power will be $(2+k) I_{ref} V_{dd}$ and any dynamic contributions will be negligible (switching activity of the comparator is contained within the bias current). Note that the sizes of M_0 and M_0 can be skewed to mirror more or less current which can decrease power and boost output frequency, but the sizing difference is generally limited by mismatch which has the potential to contribute significant temperature dependence. If the comparator is instead implemented as clocked design that is independently operated at some frequency f_{cmp}, then a dynamic component must also be considered based on that switching frequency. It is important to note here that the comparator delay t_{cmp} has a potential to add significant additional delay (and therefore temperature dependence) to the overall clock period. Delay in this case applies to both the output propagation delay of the comparator as well as the amount of time required for the capacitor to be discharged to reset the output of the comparator. Updating (38) to account for the comparator delay, a new expression for the RXO output frequency can be given as:

$$F_{out} = \frac{1}{2\left(R_{ref} C_L + t_{cmp}\right)} \tag{39}$$

Historically, clocked comparator designs have been used to keep t_{cmp} small and minimize the impact of its temperature variation. To achieve this result, the comparator switching frequency f_{cmp} is typically is equal to (or greater than) the output frequency of the RXO itself, which leads the dynamic power of the comparator to dominate the total power of the RXO. One approach to manage the comparator power is to power-gate it for a short time within the output period of the main RXO, at which point the power becomes limited by the static biasing currents and RC network [76, 86]. Continuous-time comparators can achieve a comparably small t_{cmp} if a large bias current (large k) is used. Usually, lower power consumption is targeted by scaling down the bias current as was previously mentioned, which unfortunately this leads to much more temperature variation, although it has been shown that the temperature variation of t_{cmp} can be reasonably minimized by carefully designing the temperature variation of R_{ref} to cancel it out [77]. Assuming a continuous-time comparator is used, the total power of the RXO can then be written as

$$P_{RXO} = V_{dd}(2+k)\frac{V_{in}}{R_{ref}} \tag{40}$$

Taking (40) and (38) together, it is useful to note that they have an identical scaling dependence on R_{ref}, which also plays a key factor in limiting the power consumption of the design. Making R_{ref} very large will reduce power, but begin to require a prohibitively large silicon area and additionally reduce F_{out}. F_{out} can be compensated by making C_L smaller, but the minimum bounds on the size C_L due to the impact of parasitics restricts its value from being decreased below a certain point.

4.1.2 The Frequency-Locked Loop

When clock frequencies in the kHz-MHz range are desired, the RXO becomes less efficient for two main reasons: First, a higher reference current is needed as shown by eq. (38), which increases the power consumption. Second, the switching speed of the comparator must be faster so that its delay contribution remains negligible compared to the RC time constant, which requires more bias current that again increases the power consumption of the comparator. The FLL is a relatively newer on-chip clock generation architecture that solves both of these issues, making it an attractive architecture for high-frequency clock generation.

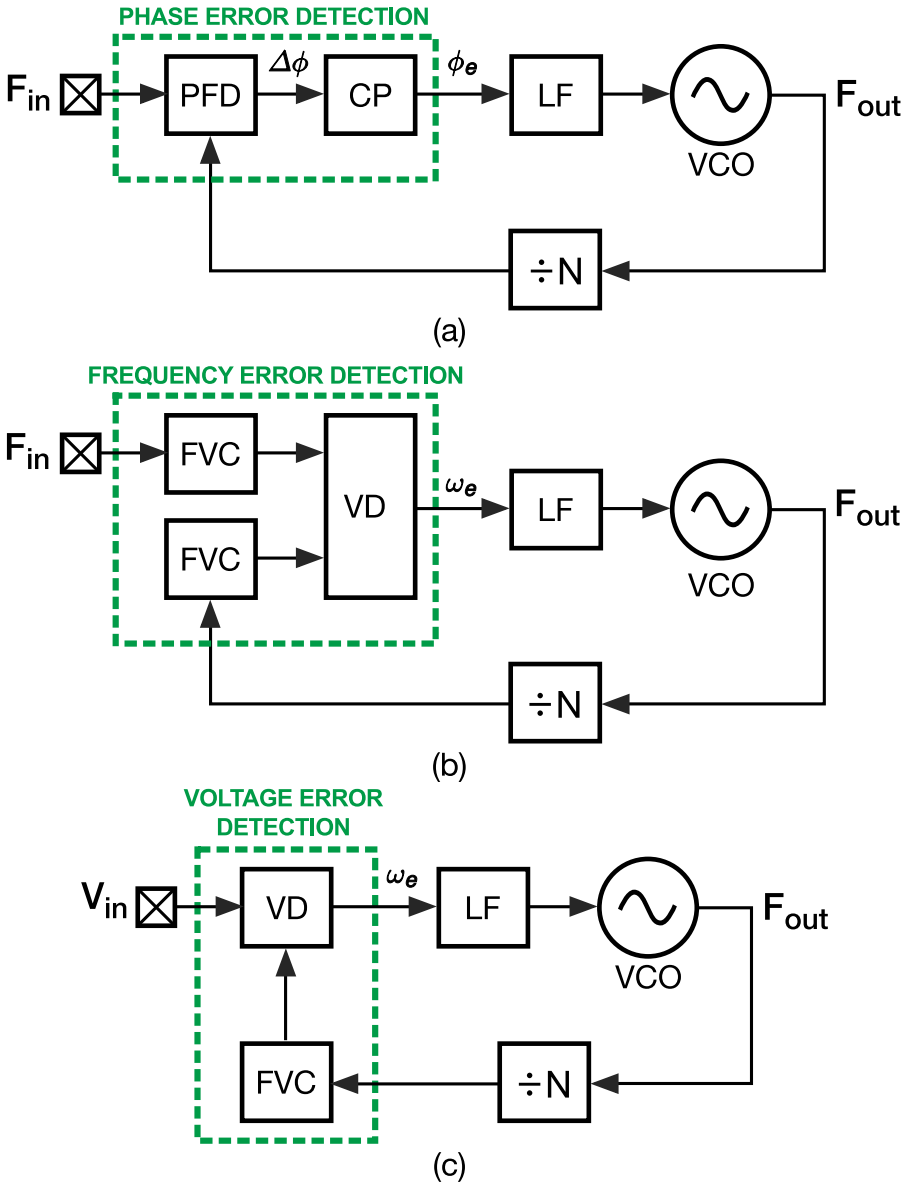

Fig. 4.3. (a) Phase-Locked Loop (PLL) architecture for frequency synthesis, (b) Frequency-Locked Loop (FLL) architecture for frequency synthesis, and (c) Frequency-Locked Loop for frequency generation.

A convenient starting point to describe the operation of the FLL is with the more commonly-known Phase-Locked Loop (PLL) pictured in Fig. 4.3(a). The PLL takes an input frequency F_{in} and generates a new output frequency $F_{\text{out}} = N \times F_{\text{in}}$ by using a phase-frequency detector (PFD) to detect the phase error Δ_ϕ between F_{in} with a divided-down copy of F_{out}, where

$$\Delta_\phi \propto \phi_{\text{in}} - \phi_{\text{out}} \tag{41}$$

Due to the feedback of this structure, F_{out} will eventually settle to a value that is in-phase (and therefore identical in frequency once divided by N) with F_{in} such that $\partial\Delta_{\phi}/\partial dt \approx 0$. A similar result can be achieved by using an FLL (Fig. 4.3b) which ignores the phase information of F_{in} and uses a Frequency-to-Voltage Converter (FVC) and Voltage Detector (VD), which is typicall just an amplifier, to detect the frequency error ω_e between F_{in} and F_{out}. Assuming both the FVCs have an identical transfer function, then

$$\omega_e \propto F_{in} - \frac{F_{out}}{N} \tag{42}$$

Since the rest of this FLL architecture works the same way as the PLL, F_{out} will be regulated via feedback from ω_e such that it will settle at a value equal to $N \times F_{in}$, however the important difference here is that the phases of F_{in} and F_{out} are nonconvergent and therefore $\partial\Delta_{\phi}/\partial dt \neq 0$. While this can have negative implications on noise performance for certain applications, the FLL has still found many uses over time. Early examples of this classic FLL design discuss several applications for signal processing [87, 88]. Various approaches for frequency error detection have been investigated that have shown several possible FVC designs [89–92].

Both of the architectures discussed so far require an existing input frequency F_{in}, which is not assumed to be present for the case of on-chip clock generation. In this case, the traditional FLL architecture can be modified by replacing the input frequency and its accompanying FVC with an input voltage V_{in}. In this manner, the FLL behaves as a VCO by detecting the voltage error between V_{in} and the output of the FVC. As a result of the feedback from this voltage error detection, F_{out} will settle (lock) at a value that depends on both V_{in} and the transfer function of the FVC. Then,

$$\omega_e \propto V_{in} - K_{FVC}\frac{F_{out}}{N} \tag{43}$$

This behavior was originally leveraged explicitly for VCO applications where V_{in} was assumed to be a varying signal [91, 93], but was later shown to be viable for on-chip frequency generation by supplying a reasonably-stable input reference voltage V_{in} and then carefully designing the FVC to both negate the dependence of $V_{IN}$and fix the dependence of F_{out} on just a single RC time constant [81, 94]. Within the last decade, the FLL has gained mainstream

attention beginning with the silicon-proven design in [78] and from several subsequent designs that reach vastly different performance points [72, 79, 83]. The FVC, shown in Fig. 4.4(a),

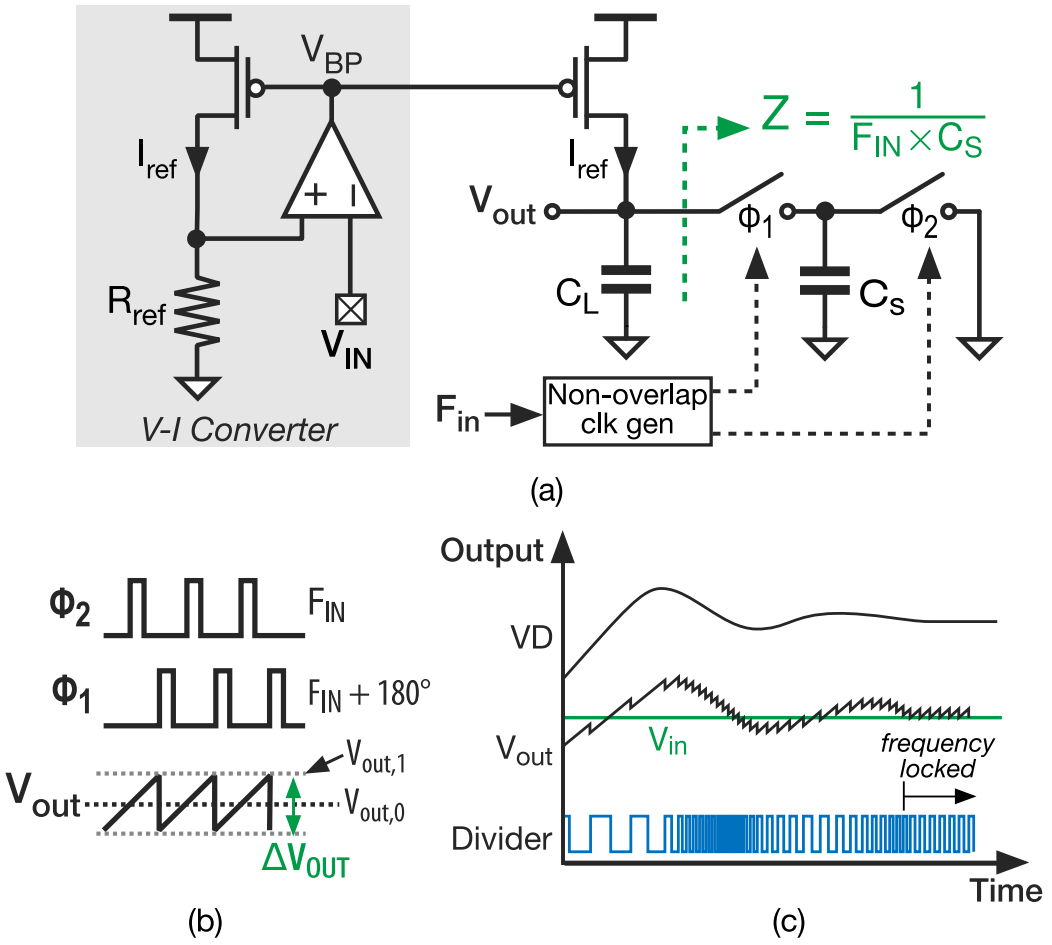

Fig. 4.4. (a) Design of Frequency-to-Voltage Converter (FVC), (b) operating waveforms of the FVC, and (c) operating waveforms for the FLL.

is similar to the RXO reference current generator shown in Fig. 4.2. It works by using a V-I converter that takes an input reference voltage V_{in} and uses a reference resistor R_{ref} to generate a reference current I_{ref}. This reference current is injected into a switched-capacitor C_S whose equivalent impedance $Z = 1/F_{in}C_S$ is notably determined by the rate F_{in} that it is switched at. Note that the switching on this node causes a small amount of output voltage ripple ΔV_{out} as shown in Fig. 4.4(b). Given an initial FVC output value $V_{out,1}$, the output ripple ΔV_{out} can be expressed as

$$\Delta V_{out} = V_{out,1} \times \frac{C_S}{C_S + C_L} \tag{44}$$

Typically, this ripple is filtered out with the Low-Pass Filter (LF) so that it does not pass to the VCO and back through the feedback loop. The resulting FVC output voltage is then solved with Ohm's Law to determine the voltage across the impedance Z when I_{ref} is injected into it:

$$V_{out} = \frac{I_{ref}}{F_{in} C_S} = \frac{V_{in}}{F_{in} R_{ref} C_S} \tag{45}$$

Since the FLL loop uses the VD to lock V_{in} with V_{out}, the resulting output frequency can be calculated as

$$F_{out} = \frac{N}{R_{ref} C_S} \tag{46}$$

Fig. 4.4(c) shows a simplified transient response of the FLL starting up and locking, where eventually the VD output settles at some DC value such that $V_{out} = V_{in}$. Note the RC time constant that is again present in this equation, but now it is scaled by the divider value N. The divider value can be adjusted on-the-fly to add frequency tunability without affecting the temperature stability (via the RC value) of the design. This has also been leveraged for on-the-fly tuning in a feedback loop to perform frequency-driven DVFS [95]. Like the RXO, the FLL will have a chunk of its power consumption dictated by the usage of bias currents (I_{ref}). The voltage detector is implemented using an op-amp that uses some factor k_1 of I_{ref} as a bias. The FVC sends I_{ref} through two branches and also uses some factor k_2 of I_{ref} to bias its amplifier. The VCO and frequency divider will add dynamic power contributions based on F_{out}. The power consumption of these components can be expressed by

$$P_{VD} = k_1 V_{dd} \frac{V_{in}}{R_{ref}} \tag{47}$$

$$P_{VCO} = F_{out} C_{vco} V_{dd}^2 \tag{48}$$

$$P_{DIV} = F_{out} C_{ff} V_{dd}^2 \sum_{i=0}^{M} \frac{1}{2^i} \tag{49}$$

$$\approx F_{out} 2 C_{ff} V_{dd}^2 \text{ for } M \gg 1 \tag{50}$$

$$P_{FVC} = (2 + k_2) V_{dd} \frac{V_{in}}{R_{ref}} \tag{51}$$

where C_{ff} is the switched gate capacitance of a single flip-flop and C_{vco} is the total switched capacitance of the VCO. The divider is assumed to use a counter-based architecture with

M bits, and its leakage power is neglected here for simplicity. The total power consumption P_{FLL} can be calculated as a sum of these components:

$$P_{FLL} = P_{VD} + P_{VCO} + P_{DIV} + P_{FVC} \tag{52}$$

$$= \underbrace{(k_1 + k_2 + 2)\, V_{dd} \frac{V_{in}}{R_{ref}}}_{\text{static power}} + \underbrace{F_{out}\,(2C_{FF} + C_{vco})\, V_{dd}^2}_{\text{dynamic power}} \tag{53}$$

The key advantage of the FLL over the RXO at high output frequencies is that R_{ref} can be made large to reduce the static power contribution from (53) without decreasing F_{out} since N can also be made large to cancel out the dependence in (46). Additionally, whereas the comparator would require a large bias current (high power) to keep its delay contribution small across temperature, the VD does not contribute a temperature-dependent delay to the overall output period and it only requires a bandwidth greater than the frequency of transient temperature fluctuations which are usually not very fast. As a result, k_1 in (47) can be made very small such that the amplifier has low bandwidth and consumes little static power. Depending on the specific value of F_{out}, the total FLL power in (53) can then be dominated either by the static power (caused by bias currents) or dynamic power (from the frequency divider and VCO). In the low-power designs prior to this dissertation, static power is typically dominant, consuming 96% of total power in [78], 83% of total power in [79], and and 70% of total power in [80]. Another issue for power consumption has been the supply voltage. Nominal supply voltages for on-chip oscillators are typically at analog-domain levels, such as 5V in 1.2μm [74] or 1.2-1.8V in 180nm [78, 79], and designs often emphasize the ability to operate across a wide range of supply voltages (e.g., output frequency should be maintained if V_{dd}drifts from 1.2V to 1.6V) which keeps dynamic power consumption high and results in energy efficiencies around 1-10pJ/cycle.

4.2 Hz-Range Relaxation Oscillator

When low-frequency clock signals are desired from an RXO, R_{ref} can simply be scaled to a high value to satisfy ((38)). The RXO is a favorable architecture in this case since the power reduction scales directly with R_{ref} and there are no other factors or components bottlenecking the power reduction. Perhaps the biggest obstacle for extending the RXO architecture to low-power and low-frequency operation is the size of the on-chip resistor, whose size can become prohibitively large in both silicon area and silicon cost. If we assume a reasonably-sized 20pF C_L is used, then a 50Hz clock signal will require a 1GΩ resistor which is not feasible to implement on-chip. One workaround for this is to use a smaller-valued resistor that results in a higher reference current, but then compensate for it by skewing the sizes of the current mirror so that only a fraction of the new reference current is actually injected onto C_L. This can successfully yield the same 50Hz frequency, but the power dissipation in the reference current generation circuit is increased due to the smaller R_{ref} . Also, as the sizing ratio of the current mirror increases, it is more difficult to avoid mismatch errors, so this technique is usually only applied on a smaller scales of current multiplication. Finally, there are other circuits and structures that can replace R_{ref} with a higher effective (nonlinear) resistance in a relatively small area, such as pseudo-resistors and diode-connected transistors, but they cannot match the same temperature stability of traditional resistors and are therefore unsuitable for use in a temperature-compensated RXO. This section discusses the use of gate leakage current (via direct tunneling through the gate oxide) as a replacement for R_{ref} with high effective resistive density and much higher temperature stability than alternative options. The reference publications for the material in this section are [DT13, DT20].

4.2.1 Gate Leakage as a Reference Resistance

This work uses a new method for generating high-valued resistances on chip without requiring a large silicon area or sacrificing a large amount of temperature stability. The technique, shown in Fig. 4.5, replaces a traditional 2-terminal R_{ref} with a transistor in a 2-terminal configuration where the source, drain, and body are tied together and act as the output node of the resistor, and the transistor gate acts as the input node of the resistor. This can be done

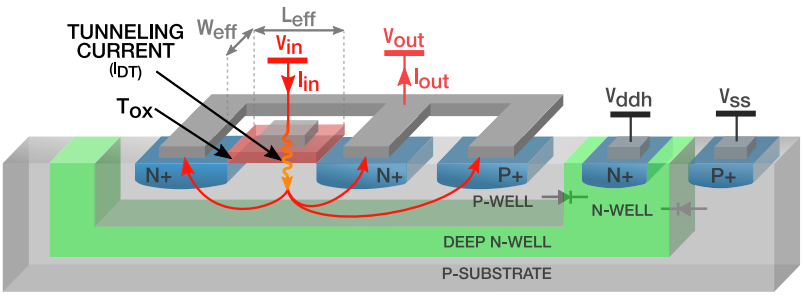

Fig. 4.5. Layout diagram of implementation for gate leakage-based resistive segment.

with either an NMOS or PMOS, but is shown here with a PMOS. Due to the limited thickness of the transistor's gate oxide, a small number of electrons will tunnel directly through the gate oxide and into the source, drain, or body of the transistor as described in 2.1.1. An even smaller amount of current will travel through the buried diodes created between p-well and deep n-well (PMOS) or between the n-well and p-substrate (NMOS) which can be modeled as in (13). Then, the total current into the positive terminal is expressed by

$$I_{out} = I_{in} + I_{diode} \tag{54}$$

The input current I_{in} is composed of multiple gate leakage components: A small amount of temperature-independent current flows directly from the gate to the source and drain (I_{gd}, I_{gs}), while a relatively larger amount of temperature-dependent current flows directly through the channel (I_{gc}) and is then gathered by the source, drain, and body contacts.

$$I_{in} = I_{gd} + I_{gs} + I_{gc} \tag{55}$$

$$= W_{eff} k_a (V_{in} - V_{out})^2 + W_{eff} L_{eff} k_b (V_{in} - V_{out}) \frac{kT}{q} \log\left(1 + e^{q(V_{in} - V_{out} - V_{th0})/kT}\right) \tag{56}$$

If the logarithmic portion of (56) is assumed to have negligible scaling across temperature, then it becomes clear that the current flowing through the structure will have a linear dependence on temperature. Note that the units of T are in Kelvin, so an increase from -20°C to 100°C would equate to a theoretical deviation of around 47% for I_{in}. If we temporarily ignore the nonlinearity of (56) by assuming that V_{in}-V_{out} can be fixed, then it can be noted that simply decreasing the width of the transistor will linearly scale down the current that travels

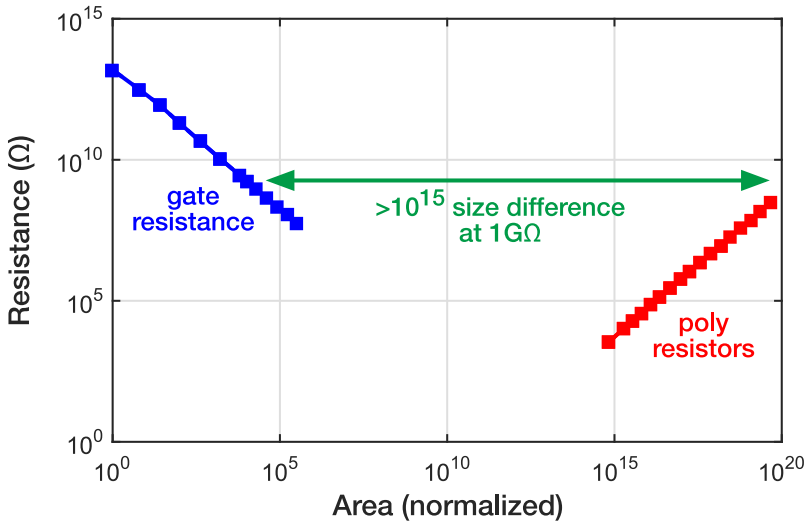

Fig. 4.6. Comparison between resistive densities of traditional on-chip polysilicon resistors and the effective resistance of the transistor gate. To create a 1GΩ effective resistance, the two structures differ in their two-dimensional silicon area by over 10^{15}x.

through the device, yielding an increase in the effective resistance R_{gate}. This is in contrast with traditional resistors that require more physical area to increase resistance as dictated by the simple equation

$$R_{\text{traditional}} = \rho \frac{L}{A} \tag{57}$$

where ρ is the resistivity of the material and L and A are the length of the segment and cross-sectional area, respectively. In on-chip resistances, the cross-sectional area is generally fixed, so creating higher value resistances requires longer segment lengths that increase the two-dimensional silicon area. Using the example from the introduction where a 1GΩ R_{ref} is desired for 50Hz clock signal, Fig. 4.6 shows the normalized area required from both a traditional on-chip polysilicon resistor and the proposed gate-leakage mechanism. For a 1GΩ effective resistance, the gate leakage technique reduces area by over 10^{15}, and provides even larger savings as the effective resistance value increases.

A challenge in implementing gate leakage as a reference resistance source is incorporating tunability which is necessary for both trimming error across chips as well as enabling either static or dynamic temperature compensation of the resistance value across temperature. When standard resistors are used, the typical approach is to break the total value of R_{ref} into N segments (each with value $R_{\text{ref}}/N\,\Omega$) that can be optionally shorted out with a bypass

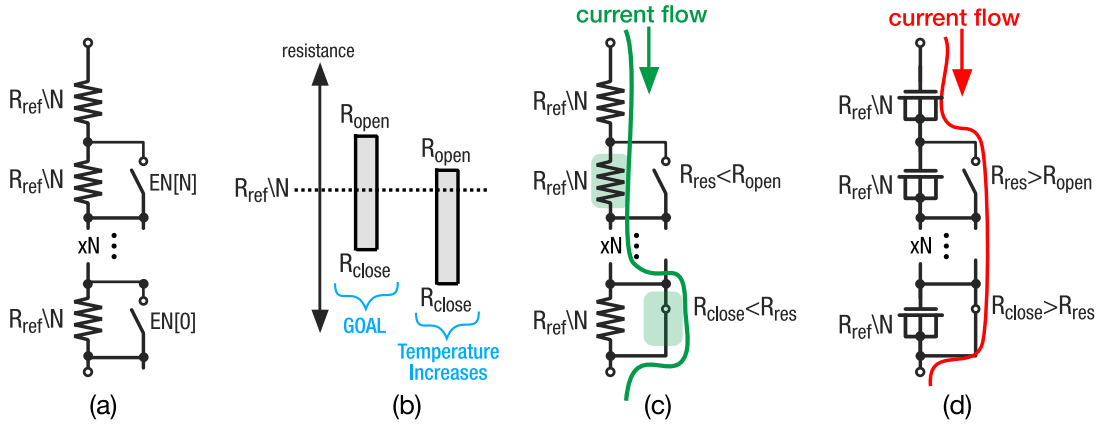

Fig. 4.7. Traditional technique for trimming an on-chip resistor (a), design goal for resistive segment bypass switches (b), implementation with standard resistors (c), and implementation with proposed gate leakage-based resistive segments (d).

switch, as shown in Fig. 4.7(a). As different segments are enabled or disabled, each resistor in the series will see a different voltage drop which is not a problem given that are ideal resistors with a linear I-V relationship. The on- and off-currents of the bypass switch should be adjusted such that the switch on-resistance R_{close} is less than R_{ref}/N such that it will adequately short out the resistive segment when activated, and the off-resistance R_{open} is high enough that no parasitic current will leak through the switch while it is inactive. This concept is shown in Fig. 4.7(b). This task is complicated by the fact that the native resistances of the switch will change across temperature. Assuming the switch can be properly designed as shown in Fig. 4.7(c), the total series resistance can be accurately configured without affecting the linearity or temperature stability of the overall structure. This approach cannot be taken for the proposed gate leakage-based resistance for two reasons. First, the nonlinearity of each segment would cause the effective series resistance to scale nonlinearly if any individual segments are enabled or disabled, which counteracts the purpose of adding this type of precision tunability. Second, the effective series resistance R_{ref}/N for a gate leakage segment is so high that it is not possible to design a bypass switch with low enough off-resistance (R_{open}) to suppress the flow of leakage current when the switch is open as shown in Fig. 4.7(d).

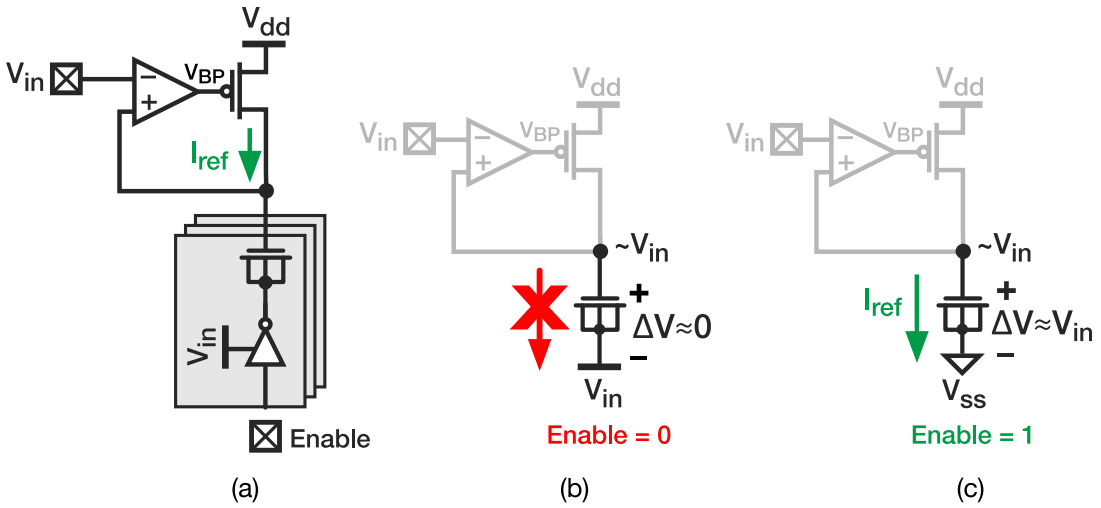

Fig. 4.8. Proposed reference current generator based on gate leakage that enables a tunable transistor width to scale the value of the reference current.

4.2.2 Gate Leakage-Powered RXO Design

To first address the need for resistance tuning, a new approach is proposed that places several gate leakage structures in parallel at the regulated note (+ input to the amplifier) as shown in Fig. 4.8(a). The bottom terminal of each structure is driven with an inverter whose supply voltage is equal to to the input reference voltage V_{in}. Then, depending on whether the inverter input (`Enable`) is logic high or logic low, the bottom terminal of a given resistive segment will be driven to either V_{in} of 0V. The the first case, when the bottom terminal is driven to V_{in}, there will be no voltage drop across the device due to the internal regulator of the top terminal of the device to V_{in} as well, so no current will flow through the device and its effective resistance will approach infinity as shown in Fig. 4.8(b). In the latter case, when the bottom terminal is pulled to 0V, a voltage drop of V_{in} will be created across the device such that some amount of current will be allowed to flow through the device, as shown in Fig. 4.8(c). Again, since the voltage at the top of the device is regulated to V_{in} through feedback, the voltage across the device will always be limited to one of these two cases, so the nonlinearity of the device is no longer an issue. Applying (56) to the structure in Fig. 4.8(a), the current through an enabled segment can be roughly approximated by assuming the logarithmic term

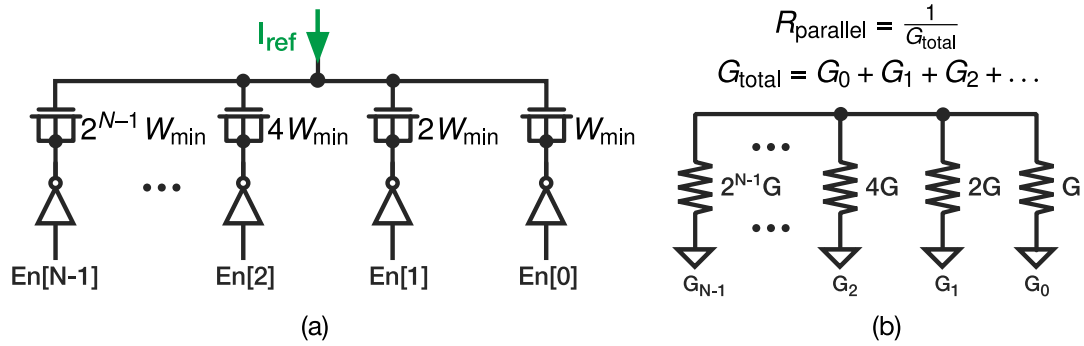

Fig. 4.9. Parallel arrangement of individual gate leakage segments to enable tunable size adjustment (a) and concept applied to traditional resistances by considering the gate leakage as a conductance G.

from I_{gc} is constant and by assuming L_{eff} is fixed such that a new constant can be created

$$k_c = L_{eff} k_b \log\left(1 + e^{q(V_{in} - V_{out} - V_{th0})/kT}\right) \tag{58}$$

Then,

$$I_{segment} \approx W_{eff}\left(k_a V_{in}^2 + k_c \frac{kT}{q} V_{in}\right) \tag{59}$$

Since I_{ref} is equal to the sum of $I_{segment}$ from all enabled structures (disabled ones don't pull any current), it becomes evident that I_{ref} can be "tuned" by enabling more or less individual segments. From (59) it is also evident that the amount of current from any given segment can be linearly adjusted by changing the width of the transistor gate. Therefore, rather than using an array of N equally-sized structures, the sizes are binary-weighted as shown in Fig. 4.9(a), such that the first segment uses a gate with $W = W_{min}$, the second segment uses a gate with $W = 2W_{min}$, and so on, until the Nth element which uses $W = 2^{N-1} W_{min}$. Then, the enable signal may be expressed as a N-bit binary value that enables an I_{ref} scaling range of $2^N - 1$. This concept can also be reached by representing each segment as a fixed conductance, as in 4.9(b), where increasing the width of any segment by some factor will increase its conductance by the same amount. Then, I_{ref} can be expressed as

$$I_{ref} \approx EN \times W_{eff}\left(k_a V_{in}^2 + k_c \frac{kT}{q} V_{in}\right) \tag{60}$$

where EN is the N-bit digital enable value that can range from 0 to $2^N - 1$.

A RXO architecture is designed to leverage the proposed I_{ref} by using two individual

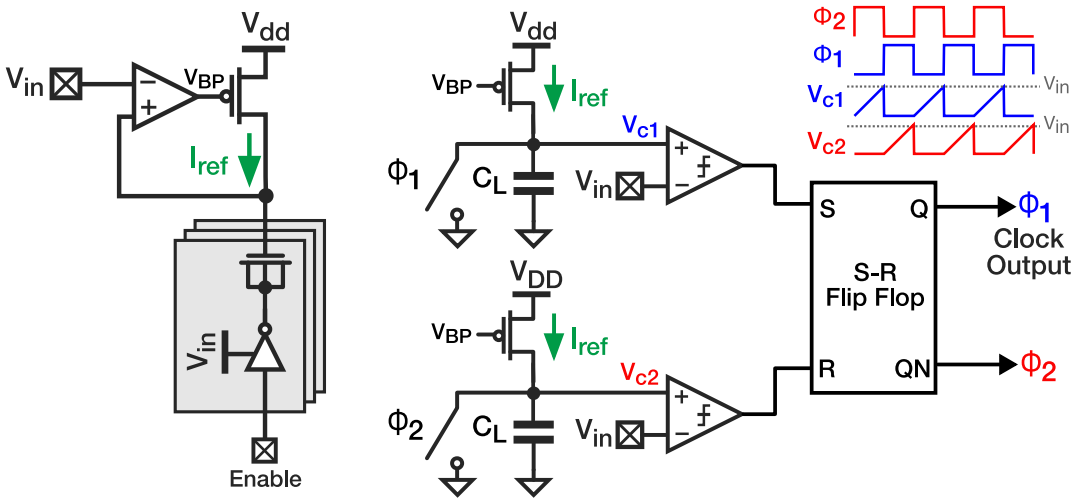

Fig. 4.10. Gate leakage-powered oscillator design with tunable I_{ref} generator and dual-phase operation.

relaxation oscillator in dual-phase operation. The design, shown in Fig. 4.10, uses the fundamental core RXO design from Fig. 4.2. As described in the derivation of (39), the additional delay from the comparator operation can skew the output frequency and add additional temperature dependence. In a standard RXO, this delay is a sum of the propagation delay of the comparator output and the time required to fully discharge C_L so that it can begin recharging on the next cycle. The comparator propagation delay can usually be made small enough to be negligible or it can be intentionally designed to have a predictable temperature sensitivity that can be leveraged for overall temperature compensation purposes. The capacitor discharge time can also be made negligible if the reset switches are strong enough to sink a high instantaneous current from C_L to ground, but this comes at the cost of increased leakage through the switch. In cases where I_{ref} is relatively high, the switch leakage is negligible and therefore validates a worthwhile design tradeoff to minimize the discharge time. However, in this work with a very small I_{ref}, the switches must be intentionally designed for very low leakage so as to not disturb I_{ref}. This means that the switches will take more time to discharge C_L and add significant extra delay to the overall output period. The dual-phase operation of the proposed design overcomes this issue by having only one RXO operating at a time. As soon as one RXO charges its capacitor to V_{in}, the output of the clock instantly toggles and the other RXO (whose capacitor is already at 0V) begins charging to time the second half of the clock period. While the second RXO charges, the first RXO has plenty

of time for its C_L to discharge. This approach does sacrifice on energy efficiency by using additional bias current, translating to a higher k in (40). Additionally, while the comparator timing delays can be addressed, there is still a temperature-dependent offset voltage V_{os} in each comparator that often matches the temperature dependence of the bias source (I_{ref}). This offset voltage causes the threshold voltage of the RXO operation to change from V_{in} to $V_{in} + V_{os}$, where V_{os} is the offset voltage of the comparator which can be expressed as

$$V_{os} = V_{os0}(1 + \alpha_{os} T) \tag{61}$$

Using (38) as the basis and assuming both parts (comparators) of the RXO are identical with no mismatch, we can equate t_{charge} as $C_L(V_{in} + V_{os})/I_{ref}$ where I_{ref} is taken from (60), and t_{charge} must also be multiplied by 2 to account for the dual-phase operation. Then, F_{out} of the RXO can be expressed as

$$F_{out} = EN \times \frac{W_{eff} V_{in}(k_c kT + k_a V_{in})}{2qC_L(V_{in} + V_{os0} + \alpha_{os} V_{os0} T)} \tag{62}$$

The specific implementation for gate leakage tuning used N=10 bits of tuning, where the first 7 bits are binary-weighted for fine adjustment and the upper 3 Most-Significant Bits (MSBs) are repeated sizes of the largest binary-weighted value for course adjustment. This 10-bit structure is then replicated 6 times, for 60 total tuning bits, where each of the 6 clusters of 10 bits uses a unique device type (NMOS, PMOS, different threshold voltage combinations). This allows exploration into the effects of different device-specific factors (k_a, k_b V_{th0} in (56)) on the temperature stability. A supply voltage of V_{dd}=1.0V is used, V_{in}=0.5V, and C_L=820fF.

4.2.3 Measurement Results

The RXO was fabricated in a low-power 65nm process test chip and consumes an area of 0.015mm^2. Fig. 4.11 shows an annotated die micrograph of the clock that is integrated together with other components as part of a larger system. This die also includes a fully-integrated temperature sensor [DT20] which will used as a dynamic compensation approach to further improve the temperature stability of the RXO. Measurements were taken from -40

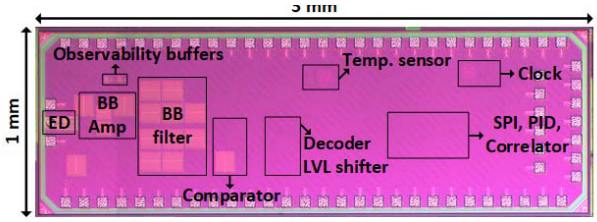

Fig. 4.11. Fabricated chip in 65-nm CMOS that contains the proposed gate-leakage powered RXO and an on-chip temperature sensor.

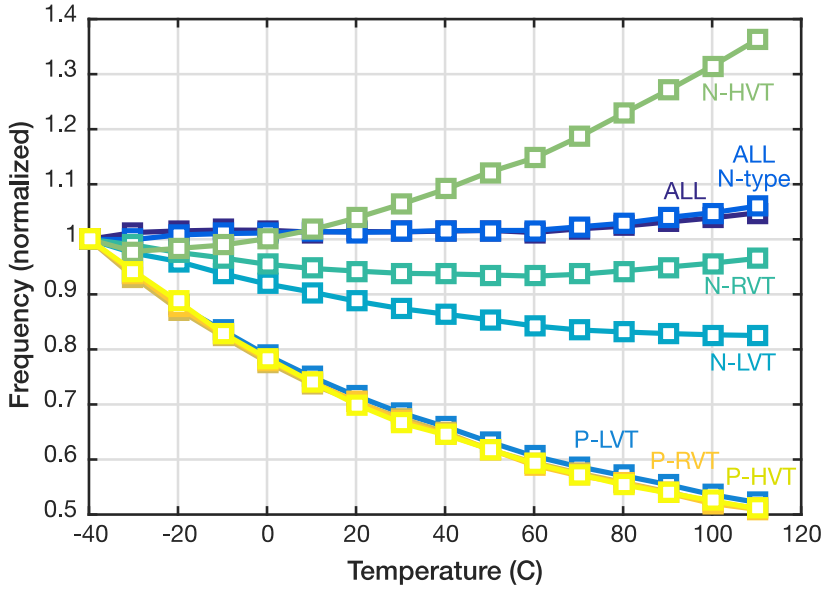

Fig. 4.12. Measured frequency of the gate leakage-powered RXO across temperature, shown for different structures that use different device types.

°C to 110°C for each individual cluster of device types (of 6 total) with all tuning bits in an individual cluster being enabled, and the results are shown in Fig. 4.12. All three PMOS devices created F_{out} values ranging from 47Hz-88Hz with negative TCs that varied by nearly 50%, or equivalent to a TC of approximately -4600 Parts-per-Million (ppm)/°C. When selecting only the LVT NMOS devices, the nominal F_{out} equaled 1.095kHz, which showed a negative TC that was less severe at a value of -1315ppm°C. The standard threshold voltage NMOS created an F_{out} of 713Hz that showed the most neutral TC of -467ppm/°C. Finally, the HVT NMOS device created an F_{out} of 272Hz that showed a positive TC of 2303ppm/°C. The fact that different device types cause different polarities of TC enables passive compensation to be used, where a mixture of devices can be enabled with both positive and negative TC in order to balance them out and get a TC as close to 0 as possible. Two examples of this are roughly

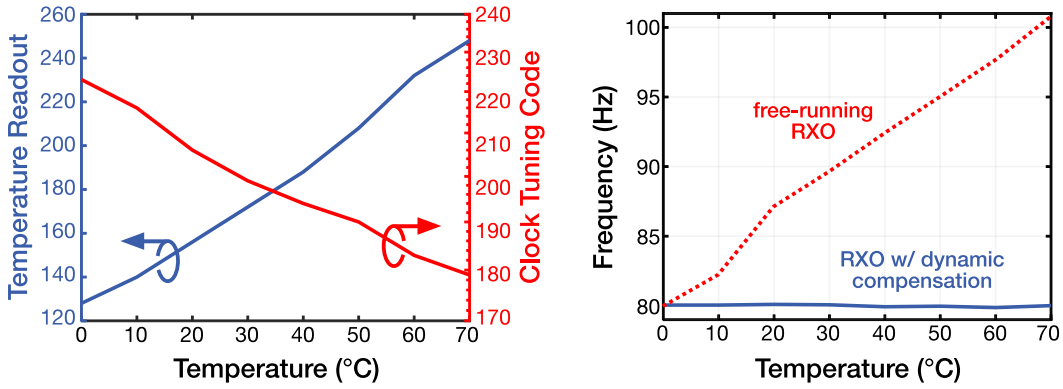

Fig. 4.13. Measurements of dynamic temperature compensation for the gate leakage-powered RXO showing the readout from the on-chip temperature sensor, the dynamically-updated I_{ref} tuning code, and the compensated and uncompensated RXO frequencies.

approximated by enabling 1) all the NMOS devices and 2) all of the devices (all 6 clusters of 10 bits). These combinations achieve TCs of 396ppm/°C and 313/°C, respectively, and create output frequencies above 1kHz. In theory, the same TC could be achieved and a lower frequency (and therefore lower power) by keeping the same ratio of enabled devices between the various clusters, but simply scaling down the total enabled values. Finally, the availability of precision tunability within any single device cluster enables dynamic compensation to be performed where an on-chip temperature sensor is used to monitor changes in temperature and a digital algorithm can change the *EN* tuning value for the RXO to keep its frequency fixed across temperature. This experiment was performed across a temperature range of 0°C to 70°C, and the results are shown in Fig. 4.13 for an 80Hz F_{out} target. The 80Hz frequency is created using a combination of NMOS HVT devices (with positive TC) to roughly set the value and then then an array of PMOS devices (with negative TC) to actuate the dynamic tuning. If the RXO is tuned to this frequency at 0°C and left running, its native frequency drift will cause an 20Hz across this range, equating to a TC of around 3100ppm/°C. Note that this value is higher than the TC of the individual HVT NMOS because it focuses only on the area of the temperature range where the F_{out} increases. The HVT NMOS frequency flattens out once the temperature drops below 0°C, which amortizes the TC to a lower value. When active dynamic compensation is added, the drift is reduced to approximately 0.5Hz, leading to a TC of 41ppm/°C. At room temperature, the RXO consumes 1.3nW while running at 80Hz, which increases to 4nW at 70°C. Table 9 summarizes the performance of the RXO for the individual gate

Structure	Max F_{out}	Tuning Resolution (Hz)	Temp Stability (ppm/°C)	Power (nW)
Individual				
N-LVT	1095	2.2	-1315	6.63
N-RVT	713	1.5	-467	4.5
N-HVT	272	1.4	2303	2.14
P-LVT	88	0.05	-4695	1.05
P-RVT	53	0.14	-4716	0.81
P-HVT	47	0.18	-4696	0.76
Combined				
All N-type	1463	N/A	396	9.1
All structures	1545	N/A	313	10
Dynamic Compensation				
N-HVT + P-HVT	80	0.18	41[a]	1.3

[a] maximum frequency error limited to 0.5Hz.

Table 9. Measurement summary of the gate leakage-powered RXO based on different compensation and tuning approaches.

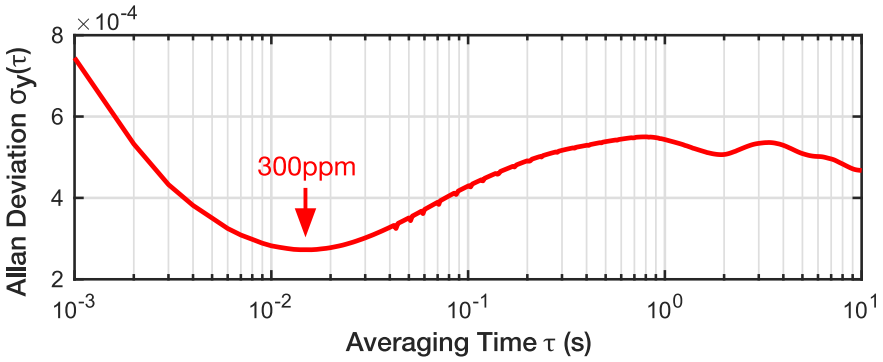

Fig. 4.14. Allan deviation measurement for the gate leakage-powered RXO showing a sub-1000ppm deviation across averaging times and a floor of 300ppm.

leakage structures and for the passive and dynamic temperature compensation approaches.

Fig. 4.14 shows the measured Allan deviation of the RXO while operating at a frequency of just under 1kHz, showing a floor of 300ppm. This calculation uses the overlapping Allan deviation approach, explained in detail by [96] but listed below here for convenience

$$\sigma_y^2 = \frac{1}{2(N-2m)\,\tau^2} \sum_{i=1}^{N-2m} [x_{i+2m} - 2x_{i+m} + x_i]^2 \tag{63}$$

where x_i is the i_{th} of N=M+1 phase error measurements taken during an averaging time of $\tau = m\tau_0$ between the oscillator of interest and its nominal frequency signal. It is interesting to note here that in contrast with resistor-based RXOs whose short term stability is typically limited by (resistor) thermal noise, the noise mechanics of gate leakage current have been

found to match shot and flicker noise [97], which can in turn give rise to different noise behavior in this type of oscillator such as higher stability at short averaging times.

4.2.4 Conclusion

This section presented a classic RXO architecture that replaces the traditional reference resistor R_{ref} with the effective resistance of a transistor's gate oxide to obtain a high resistive density for low silicon area. An analysis of the temperature stability of this reference resistance source is performed to show that a reasonably linear frequency versus temperature can be created depending on a combination of the technology-related constants for device parameters and the offset voltage of a continuous-time comparator. Design modifications are proposed to enable precision frequency tuning and temperature compensation, and a test chip was fabricated in 65-nm CMOS. Measurement results show that an intrinsic temperature stability of several hundred ppm/°C with an ability to reach sub-nW operaton in at Hz-range frequencies. When combined with a temperature sensor, the design can be dynamically tuned to reach a TC of 41ppm/°C.

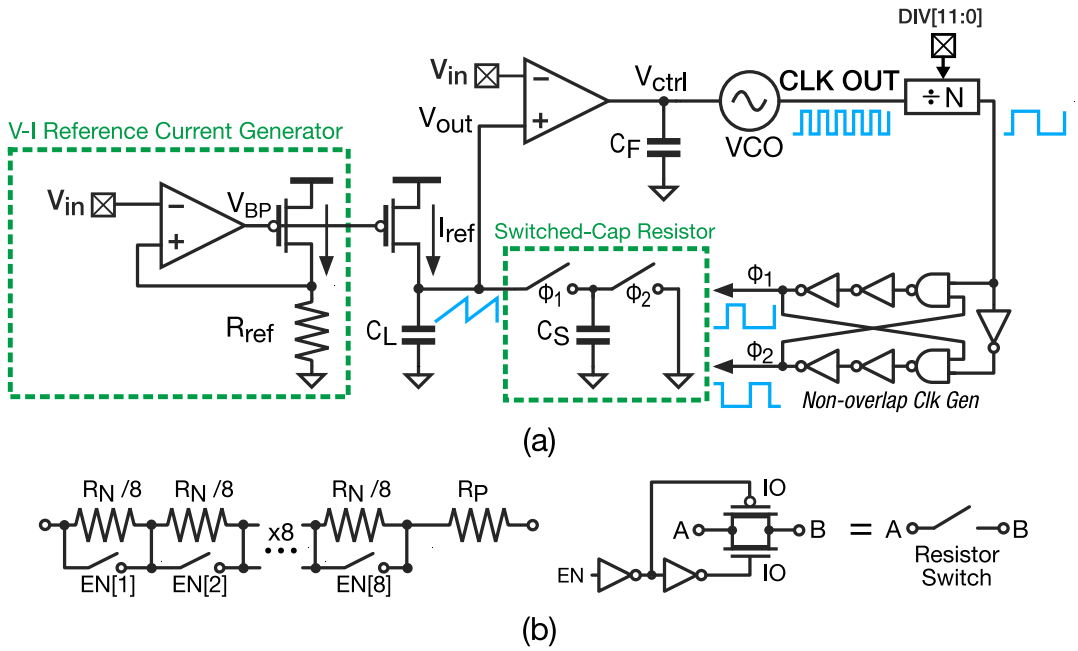

Fig. 4.15. Design of the analog FLL (a) and implementation of R_{ref} to enable passive temperature compensation (b)

4.3 Analog FLL

This section discusses the analysis, design, and performance of an energy-efficient analog FLL design that significantly reduces energy-per-cycle to a level that is over an order of magnitude lower than that of prior works while maintaining state-of-art temperature stability. To accomplish this, the analysis in 4.1.2 is continued to find an expression for the energy efficiency of the proposed FLL architecture and to introduce design optimizations to reduce the energy contributions of the biasing circuitry that typically limits FLL energy efficiency. Key design choices to enable this optimization are 1) the use of a loop divider to boost output frequency without needing to increase bias currents or static power, 2) designing the FLL to operate at low supply voltage, and 3) using an ultra-low 0.1V loop reference voltage. The reference publication for the material in this section is [DT15].

4.3.1 Design and Analysis

Fig. 4.15(a) shows the proposed analog FLL design. The FVC is implemented as in Fig. 4.4 where a V-I reference current converter uses V_{in} along with R_{ref} to create a reference current I_{ref} that is injected into a switched-capacitor resistor $C_S = 20fF$. Load capacitor $C_L = 20pF$ is

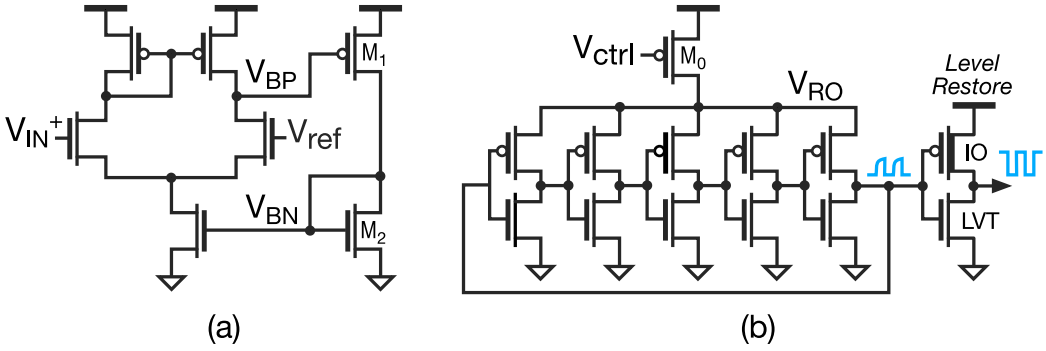

Fig. 4.16. Schematics of the analog FLL loop amplifier (A) and VCO (b).

used to minimize the ripple on V_{out} (see (44)). The output of the FVC (V_{out}) is fed into the positive terminal of an op-amp which serves as the VD. The output of the op-amp is stabilized with a filtering capacitor $C_F = 10pF$ and then used to drive the VCO, whose frequency is divided down and separated into non-overlapping signals that drive the FVC. The combination of the VD op-amp and C_L form the LF that is seen in Fig. 4.3(c). The frequency divider is made tunable by the integer value N. Fig. 4.16(a) shows the schematic diagram of the VD op-amp, whose output V_{BP} is used to internally mirror I_{REF} back as the tail bias current using transistors M_1 and M_2. The VCO is implemented as a 5-stage ring oscillator that uses a PMOS current-starver (transistor M_0) to regulate the internal RO voltage V_{ro}. Since the RO output swing will then be limited to V_{RO}, a skewed-strength inverter is used to restore the output level. The energy per clock cycle of this VCO can be expressed by

$$E_{VCO} = 5C_{inv}V_{dd}V_{ro} \tag{64}$$

As shown in the ideal F_{out} equation in (46), the simplified output frequency for this design is equal to $N/R_{ref}C_S$. On-chip capacitors generally have very low temperature dependence, but on-chip resistors can easily vary by several hundred ppm per degree Celsius. This design takes the typical approach to compensating the temperature dependence of the on-chip resistor by using two resistors in series with positive (R_P) and negative (R_N) temperature

coefficients as shown in Fig. 4.15(b). The values of these resistors can be expressed by

$$R_P = R_{P0}(1 + \alpha_{P0}\Delta T) \tag{65}$$

$$R_N = R_{N0}(1 + \alpha_{N0}\Delta T) \tag{66}$$

where R_{P0} and R_{N0} are the nominal resistances of R_P and R_N, respectively, which have temperature coefficients α_{P0} and α_{N0}, and ΔT is the deviation from the nominal temperature. These first order models do not capture the higher-order temperature dependencies that are usually present in on-chip resistors, but are accurate enough for modeling approaches such as the one in this dissertation. Then, the series resistance of R_{ref} will be

$$R_{ref} = R_{P0} + R_{N0} + (R_{P0}\alpha_{P0} + R_{N0}\alpha_{N0})\Delta T \tag{67}$$

Carefully balancing the relative sizes of R_P and R_N will enable the effects of temperature to be minimized as shown by (67). This is enabled in silicon by picking a static value for R_P and then breaking R_N into several small segments as shown in Fig. 4.15(b) that can be individually enabled or disabled based on the natural variation between design values and fabricated values. For the implementation here, the total series value of R_{ref} is approximately 49Ω. If we also assume that the amplifier in the V-I converter has some offset voltage $V_{os,1}$ (which can have its own temperature dependence), then the value of I_{ref} now becomes

$$I_{ref} = \frac{V_{in} + V_{os,1}}{R_{P0} + R_{N0} + (R_{P0}\alpha_{P0} + R_{N0}\alpha_{N0})\Delta T} \tag{68}$$

Finally, if we also assume that the loop amplifier VD has its own independent offset voltage $V_{os,2}$, then the output frequency of this design can be expressed as

$$F_{out} = \frac{N\left(V_{in} + V_{os,1}\right)}{C_S\left(V_{in} + V_{os,2}\right)} \times \frac{1}{(R_{P0} + R_{N0} + (R_{P0}\alpha_{P0} + R_{N0}\alpha_{N0})\Delta T)} \tag{69}$$

Besides ensuring that the overall TC of R_{ref} is minimized as much as possible, the only other major step required to stabilize the temperature sensitivity of this design is to ensure that the temperature dependences of $V_{os,1}$ and $V_{os,2}$ match – it does not actually matter if there is a

large offset voltage or if that offset voltage is sensitive to temperature as long as it affects both amplifiers equally, since the effects cancel out in (69). Turning the focus now toward the energy efficiency, the total energy per clock cycle of this design can be expressed by combining (64) with (53) and (46):

$$E_{\text{cycle}} = \frac{mC_{\text{S}}V_{\text{in}}V_{\text{dd}}}{N} + (2C_{\text{ff}} + 5C_{\text{inv}})V_{\text{dd}}^2 \tag{70}$$

where m represents the sum of the scaling components such as k_1 and k_2 from (53). Note that (70) takes a pessimistic modeling approach by setting the VCO energy dependence on supply voltage to V_{dd}^2 rather than $V_{\text{dd}}V_{\text{ro}}$ as in (64). Based on the expression in (70), several factors are considered in this work to design the energy-efficient analog FLL. First, the static power contributions of the VD and FVC should be minimized, so a reasonably large R_{ref} is used that does not consume excessive silicon area. Increasing the divider value N can compensate for the larger R_{ref} to maintain a high F_{out}. Additionally, a low V_{in} of 0.1V is used to further reduce the energy contributions from the VD and FVC. After these reductions to the static power sources, the dynamic power contributions from the divider and VCO need to be addressed. Since these are both quadratically dependent on V_{dd}, a reduced supply voltage of 0.6V is used for the FLL. This aggressive V_{dd} scaling does come at the cost of an inability to maintain robust operation across the range of supply voltages of many previous works. While this FLL design will continue to operate if its V_{dd} is increased or decreased beyond the nominal value, the nominal frequency will be subject to relatively high variation that also sees a worse temperature stability across the temperature range. Therefore, for this design to be fully leveraged in a system for extreme energy reduction, it is assumed that a low-noise supply voltage regulator will be available at or near the nominal supply voltage value. The VCO capacitance C_{VCO} is kept as small as possible by using a ring oscillator with the minimum possible number of stages and minimum-sized devices, and the linear regulation approach used to control the RO (using transistor M_0) reduces its energy dependency on supply voltage from V_{dd}^2 to $V_{\text{dd}}V_{\text{ro}}$, where $V_{\text{ro}} < V_{\text{dd}}$. Fig. 4.17 shows the simulated energy per cycle of the presented analog FLL along with the modeled values from (70). These results are shown versus the divider value N, which is linearly proportional to the output frequency.

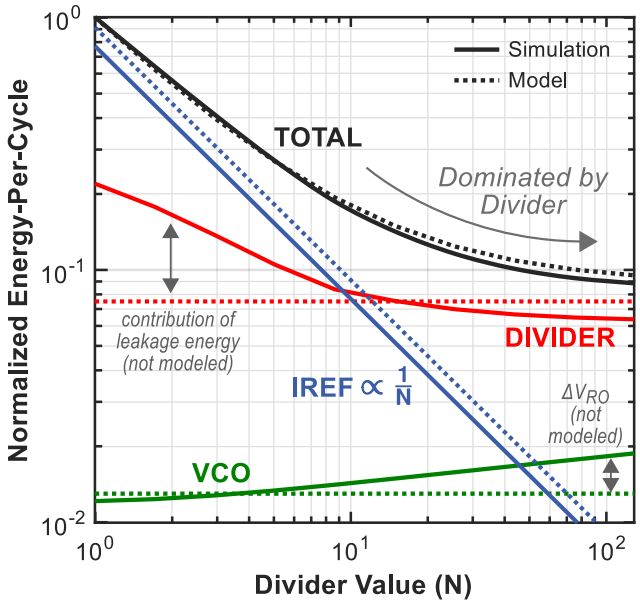

Fig. 4.17. Modeled and simulated energy consumption of the analog FLL, showing breakdown of energy between the divider, VCO, and I_{ref} Models used are from (70).

At low frequencies when N=1, static power dominates the total energy. Increasing N provides a proportionally higher F_{out} which proportionally increases P_{DIV} and P_{VCO} but causes no change in P_{FVC}. As a result, the energy contributions from the divider and VCO remain unchanged versus N while the energy contribution from the FVC decreases inversely with N. Nonidealities from the simplified modeling approach arise when F_{out} is very low which causes the leakage power of the divider (which was neglected) to become much more significant, and at high F_{out} when V_{ro} starts to deviate.

4.3.2 Measurement Results

The proposed FLL design was fabricated in a 65nm low-power process occupying an area of 0.098mm^2. Fig. 4.18 shows an annotated chip micrograph. At room temperature (23° C), the base frequency (N=1) is approximately 63.5kHz, which deviates from the theoretical value of $1/R_{REF}C_S$ by 10khz, which is equivalent to an added parasitic capacitance on C_S of around 35fF. At this operating point, it consumes roughly 25nW. Due to the limited V_{dd}, the maximum divider value that the design can support is N=16, which equates to an F_{out} of just over 1MHz. Higher values can be supported at room temperature (23°C) but not sustained at hotter or colder temperatures. Fig. 4.19 shows the measured start-up response of the FLL at

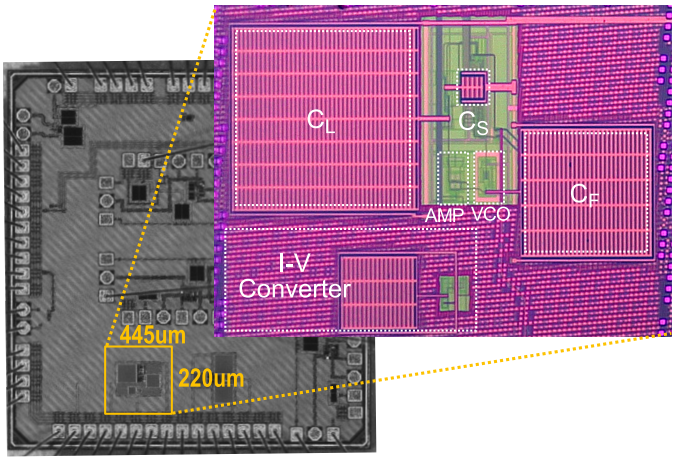

Fig. 4.18. Annotated micrograph of the fabricated analog FLL design in 65nm CMOS.

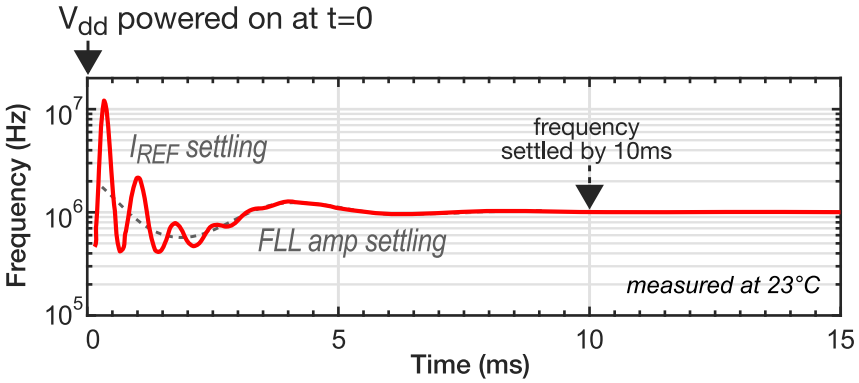

Fig. 4.19. Measured start-up response of the FLL, showing a 10ms settling time. Dashed line shows visualization of FLL amplifier settling without the added modulation from the I_{ref} startup.

the maximum divider value of N=16, demonstrating a 10ms settling time at a frequency of 1.016MHz. The start-up behavior reveals the settling of the V-I converter that stabilizes within 5ms and is modulated by the start-up response of the FLL loop amplifier, which takes another 5ms to settle. Note that [78] provides more theoretical detail on the frequency response based on a similar architecture. Fig. 4.20(a) shows the measured power and energy consumption of the FLL versus its output frequency range. As mentioned before, the output frequency can be scaled from 50kHz to 1MHz by changing the divider value. As expected, the total power scales linearly with frequency due to the dependence of (53) on the dynamic power of the VCO and frequency divider. At 580kHz, the design consumes 34nW. Dividing the power by F_{out} yields the energy per cycle of the design, which scales from 200fJ/cycle at N=1 to 44.6fJ/cycle at N=16. Again, this result closely matches the expected value from the

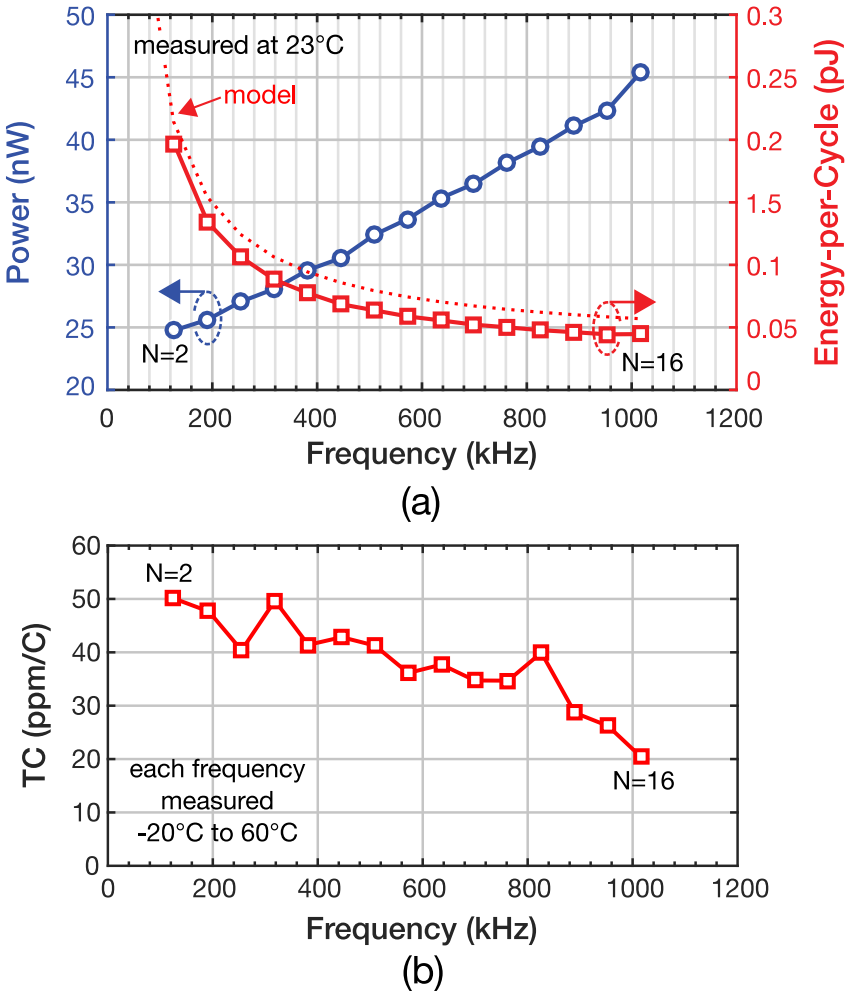

Fig. 4.20. Measured power and energy consumption of the FLL versus the output frequency, obtained by changing the divider value from *N*=2 to *N*=16 (a) and the TC of each output frequency measured from -20°C to 60°C (b).

model, which is plotted by a dashed line. The decreasing energy contribution from static bias currents as N increases causes the total energy to plateau at higher frequencies where it is fundamentally limited by the energy of the VCO and divider.

Fig. 4.20(b) shows the measured temperature stability of the output frequency for every divider value, represented by the temperature coefficient in terms of ppm per degree Celsius (ppm/°C). This result is obtained by setting the FLL divider value a specific value and then sweeping the temperature from -20°C to 60°C while measuring the output frequency in 10°C increments. This process is then repeated for all integer divider values from N=2 to N=16. All possible output frequencies maintain a stability of at least 50ppm/°C across the temperature

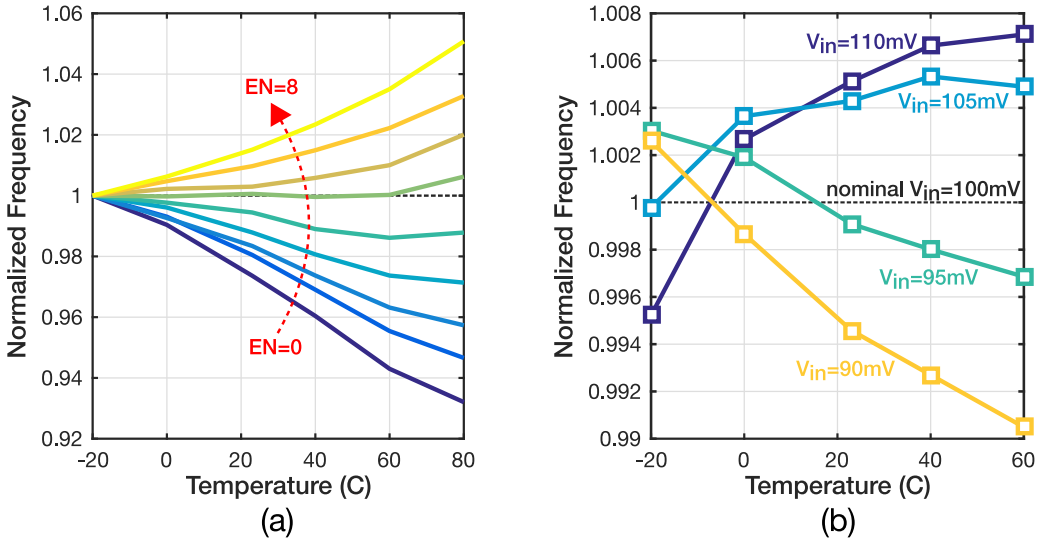

Fig. 4.21. Measured temperature compensation ability of R_{ref} shown by measuring F_{out} over the full temperature range for each discrete R_{ref} trim setting (a) and sensitivity of F_{out} to the input reference voltage V_{in}.

range, with the highest stability being achieved at N=16 with a ppm/°C of 20.3.

Fig. 4.21(a) shows the measured temperature dependence of the FLL output frequency on the R_{ref} trim setting and on the input reference voltage V_{in}. When the resistor `EN` tuning word is set to 0, all of the negative temperature coefficient resistances (R_N) are disabled, so R_{ref} will approximately equal R_P and therefore increase in value along with temperature. Since R_{ref} appears in the denominator of (46), F_{out} would be expected to decrease as temperature increases, which is indeed reflected by this measurement. Gradually increasing the `EN` tuning word increases the number of R_N segments that are enabled which begins to offset the R_P temperature coefficient as shown by (69). At the maximum `EN` value, F_{out} increases linearly across temperature. To find the optimal trim setting for R_N to stabilize F_{out} across the temperature range, F_{out} should be measured across all trim settings at two different temperatures (e.g., room temperature and 40°C) in order to find the setting that yields the smallest slope between the two temperatures. This process is known as a 2-point trim. Depending on the variation between chips and the number of tuning settings used, it should be feasible to find the average optimal setting for several chips and apply it to the rest of the chips in a batch. Note also that while this approach stabilizes the F_{out} TC, it will subsequently affect the nominal F_{out} value. If it is desired for all chips in a batch to have equal nominal frequencies, absolute frequency trimming could be implemented by adding trim to

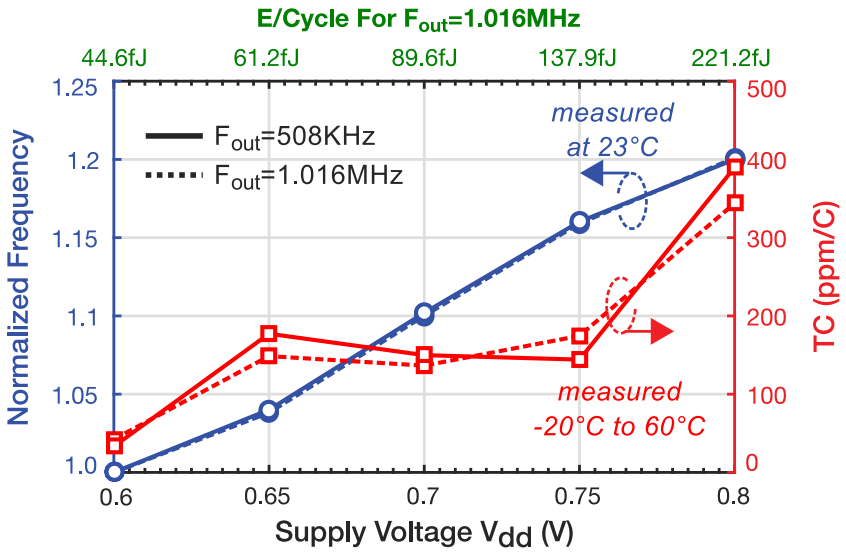

Fig. 4.22. Sensitivity of the FLL frequency and temperature stability to supply voltage, measured from the nominal 0.6V supply up to a 0.8V supply. Higher supply voltages increase the output frequency and reduce its temperature stability (higher TC). This effect impacts the FLL equally regardless of what output frequency (divider N) it is running at.

R_P or C_S. Alternatively, if only nominal frequency matching is desired, a 1-point trim can be used to tune R_N for all chips at a single temperature. While the value of V_{in} does affect the energy efficiency, it theoretically has no impact on the value or TC of F_{out} except under the condition that input-dependent offset voltage mismatch occurs between the loop amplifier (VD) and V-I amplifier as shown in (69). This point was experimentally tested by measuring the FLL output frequency across the temperature range for V_{in} values spanning $\pm$10mV from the nominal value. The results, shown in Fig. 4.21(b), show a slight dependence on V_{in}. The worst-case variation shown in this plot occurs only at 60°C if V_{in} were to change from its maximum to minimum value, which would equate to 370ppm/°C of variation. However, this result is not realistic as the value of V_{in} would actually be changing across temperature and not within a temperature, so the resulting F_{out} TC would be lower than this given value depending on the actual temperature characteristic of V_{in}. Restricting V_{in} to a worst-case variation of just 2mV would render its impact on F_{out} negligible, and would only impose a TC requirement of 250ppm/°C on V_{in} across the temperature range.

Fig. 4.22 shows the effect of supply voltage variation on the value and TC of F_{out}, both of which incur significant variation due to the subthreshold biasing of the FLL amplifiers

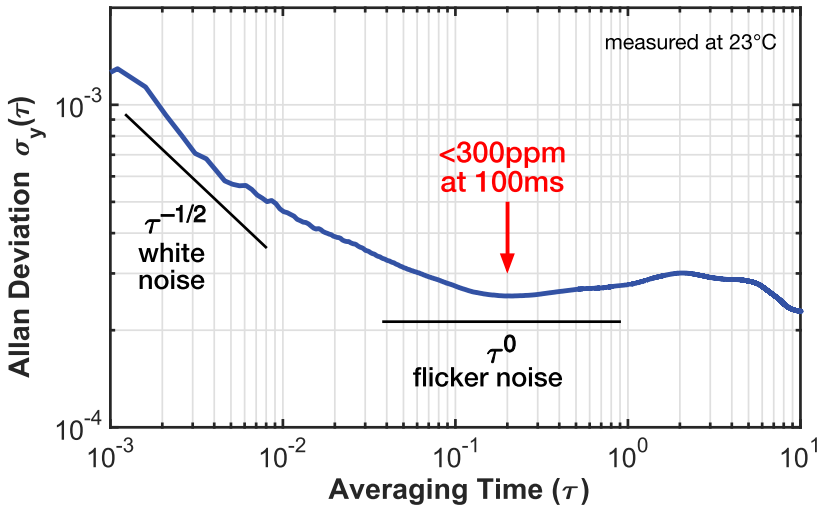

Fig. 4.23. Allan deviation measurement of the analog FLL showing a flicker noise-limited floor of 300pm at an averaging time of 100ms.

(average of 111%/V supply sensitivity across 18 dies). Increasing V_{dd} beyond the nominal value of 0.6V up to 0.8V increases F_{out} by up to 20% and reduces the temperature stability to several hundred ppm°C. These effects are nearly identical at all possible output frequencies. As mentioned previously, the expectation for this design is that it will run from a regulated supply voltage of 0.6V, which is a typical value present in IoT systems since it can be used for low-power analog circuits as well as near-threshold digital circuits that also target energy efficient operation. Having a stable supply voltage is also critical for the FLL to maintain low energy consumption, as higher supply voltages would also increase the energy per cycle as shown on the top x-axis of Fig. 4.22.

Finally, the time-domain stability of the analog FLL is assessed by calculating its Allan deviation for an averaging time of up to 10 seconds. The results, shown in Fig. 4.23 indicate that the design reaches a variance of less than 300ppm for an averaging time beyond 100ms. Table 11 summarizes the key performance metrics of this analog FLL design presented in this section and compares its performance to other state-of-the-art works at the time of publication.

	This Work	[DT18]	[98]	[79]	[99]	[78]	[75]
Technology (nm)	65	65	40	180	180	180	180
Area (mm^2)	0.098	0.051	0.07	0.16	0.5	0.26	0.03
Supply Voltage (V)	0.6–0.8	1.0	0.65–0.8	1.0–1.8	0.85–1.4	1.2–1.8	0.6–1.8
Frequency (kHz)	1016	1050	417	32.7	3.0	70.4	122
Temp. Range (°C)	-20–60	0–40	-20–80	-20–100	-25–85	-40–80	-20–100
Temp. Stability (ppm/°C)	20.3	2.5	106	13.2	13.8	34.3	327
Power (nW)	45.3	69000	181	35.4	4.7	110	14.4
Energy per Cycle (pJ)	0.0446	65.7	0.43	1.08	1.6	1.56	0.12

Table 10. Performance summary of analog FLL design and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (10/10/19), and more recent works are not shown here. Reference publication for this work is [DT15].

4.3.3 Conclusion

This section presented an analog FLL architecture and performed a theoretical analysis of its energy efficiency and temperature stability. The derived models are used to implement an energy-efficient analog FLL that is fabricated in 65-nm CMOS. Measurement results show that the analog FLL achieves the best energy efficiency at the time of publication (44.6fJ/cycle) and operates over a wide range of frequencies (60kHz to 1MHz) while maintaining a high temperature stability that peaks at 20.3ppm°C.

4.4 Duty-Cycled Digital FLL

As shown in Section 4.3, the power consumption of the analog FLL can be reduced to a point where static power power contributions are negligible and only the dynamic power of the divider and VCO are dominant. This section introduces a DFLL architecture that duty-cycles the divider and biasing circuits to reduce their power further. This is achieved by "freezing" the VCO frequency once the FLL has has settled so that the the divider and biasing circuits can be temporarily disabled without causing the loop to become unstable from the disrupted feedback. Running the FLL in this open-loop configuration leaves the VCO susceptible to temperature variation since it does not have good intrinsic temperature stability. To compensate for temperature drift, the divider and biasing circuits are periodically re-activated so that the FLL can re-lock. The reference publication for the material in this section is [DT2].

4.4.1 Digital FLL Design

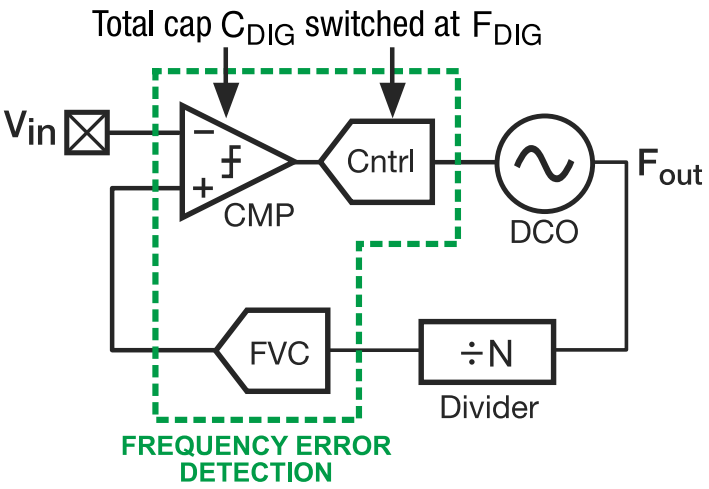

Fig. 4.24. Generalized architecture of a digital FLL.

To address the design goals for this work, a digital FLL implementation is chosen which has a key advantage over the analog FLL which will be discussed later. The digital implementation differs from the analog implementation in that the VD digitizes the frequency error and uses this digital value to drive a DCO instead of a VCO. This is performed by using a comparator (rather than an amplifier) to sample the voltage difference between V_{in} and V_{out}

and a control ("Cntrl") block such as a Successive-Approximation Register (SAR) to tune a digital value that controls the frequency of the DCO. While it may seem counterintuitive to add a comparator back into the architecture since its absence is what originally gives the FLL an energy-efficiency advantage, this comparator actually does not need to be clocked very fast and therefore dissipates a very small amount of dynamic power.

The DFLL power consumption is still modeled by summing the individual contributions from each component as in (52), but the expression for the VD must be updated to reflect the digital circuits that contribute both dynamic and leakage power. A new expression is created here for convenience that represents the power of both the VD and the FVC that together constitute the "locking" circuitry:

$$P_{\text{lock}} = P_{\text{VD}} + P_{\text{FVC}} \tag{71}$$

$$= mV_{\text{dd}}\frac{V_{\text{in}}}{R_{\text{ref}}} + F_{\text{dig}}C_{\text{dig}}V_{\text{dd}}^2 + P_{\text{leak}} \tag{72}$$

where again m is used as a scaling factor to represent the total usage of I_{ref}, which in this case is equal to $m = 2 + k_2$ as in (51), while k_1 is excluded due to the VD (clocked comparator) not requiring a k_1 multiple of I_{ref}. The comparator and digital controller contain a total effective switched-capacitance C_{dig} that operates at an independent frequency $F_{\text{dig}} \ll F_{\text{out}}$ that keeps their dynamic power low. Their leakage power P_{leak} can be modeled as $NI_{\text{leak}}V_{\text{dd}}$ where N is the number of transistors in the design and I_{leak} is the leakage current of a single transistor at V_{dd}. Using this expression with (52), the energy per cycle of the DFLL can be written as a function of F_{out}

$$E_{\text{cycle}}(F_{\text{out}}) = (2C_{\text{ff}} + 5C_{\text{inv}})V_{\text{dd}}^2 + \frac{P_{\text{lock}}}{F_{\text{out}}} \tag{73}$$

Replacing the final term in (73) with (72) and (46) yields a more complete model that can be expressed instead as a function of the divider N

$$E_{\text{cycle}}(N) = (2C_{\text{ff}} + 5C_{\text{inv}})V_{\text{DD}}^2 + \frac{C_{\text{S}}R_{\text{ref}}}{N}\left(NI_{\text{leak}}V_{\text{dd}} + C_{\text{dig}}F_{\text{dig}}V_{\text{dd}}^2 + \frac{mV_{\text{dd}}V_{\text{in}}}{R_{\text{ref}}}\right) \tag{74}$$

At baseline, the energy efficiency of this DFLL design is comparable to the analog FLL as modeled by (70). While P_{VD} can be somewhat lower due to the reduced m value, this

architecture still has the potential to be limited by either dynamic or static power depending on F_{out}. Designs with 1) high bias currents or auxiliary amplifiers (large m), 2) complex digital control algorithms (high N, C_{dig}, or F_{dig}), or 3) no loop divider have relatively higher static energy that tends to push total energy consumption to the pJ/cycle range. Decreasing V_{dd} will reduce energy but increase the relative contribution of the static energy since its near-linear dependence on V_{dd} will be out-scaled by the quadratic dependence that the VCO/DCO and divider have.

4.4.2 Steady-State Frequency Inaccuracy

The output frequency of the DFLL is subject to nonidealities and variation from several components. Temperature-dependent inaccuracy comes from variation in R_{ref} and the offset voltages of the amplifier in the V-I converter and the comparator in the voltage detector, similar to the Analog FLL. Additionally, the FVC output voltage ripple ΔV_{out} shown in (44) limits the voltage error detection ability which results in a finite steady-state frequency locking error. Finally, the DCO gain and non-linearity can can further deteriorate the locking resolution. Temperature-dependent inaccuracy is accounted for here the same way as in the analog FLL, where a new expression is created for F_{out} that assumes the voltage inputs to the amplifier and comparator are not equal:

$$F_{out} = \frac{NV_{in,amp}}{R_{ref}C_S V_{in,cmp}} \tag{75}$$

where $V_{in,amp}$ and $V_{in,cmp}$ account for the different temperature-dependent offset voltages of the amplifier and comparator, respectively. All temperature dependencies are reflected by the same first-order models that were previously used for the analog FLL:

$$V_{in,amp} = V_{in0}\left(1 + \alpha_{amp}\Delta T\right) \tag{76}$$

$$V_{in,cmp} = V_{in0}\left(1 + \alpha_{cmp}\Delta T\right) \tag{77}$$

$$R_{ref} = R_0\left(1 + \alpha_R \Delta T\right) \tag{78}$$

where the TCs of the comparator (α_{cmp}), amplifier (α_{amp}), and R_{ref} (α_R) are expressed in units of 1/°C and ΔT is the deviation from a nominal temperature T_0. Temperature variation

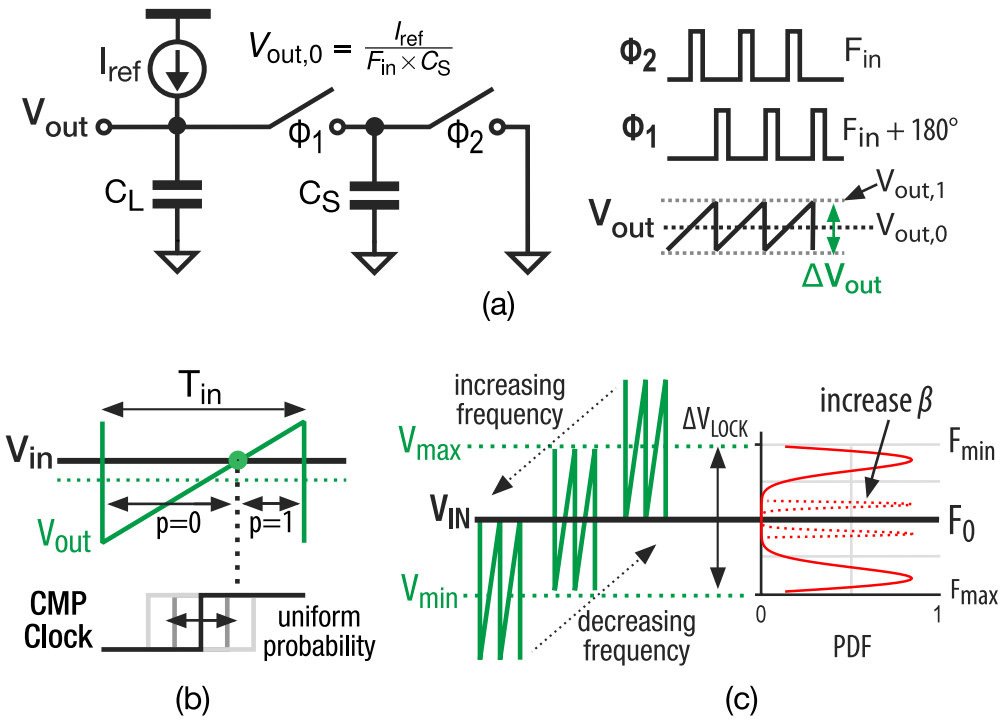

Fig. 4.25. Simplified schematic and timing diagram of the FVC (a), illustration of locking probability during a decreasing frequency search (b), and resulting voltage/frequency locking range (c).

in V_{in} can be accounted for in this model by adding the TC of V_{in} to both α_{amp} and α_{cmp}. Fig. 4.25 illustrates how the FVC output voltage ripple ΔV_{out} leads to a voltage detection error that limits the steady-state frequency locking range. Given an initial FVC output value $V_{out,1}$, the output ripple ΔV_{out} expressed by (44) is shown in Fig. 4.25(a). In an analog FLL, the effect of the output ripple on F_{out} depends on the gain and bandwidth of the loop amplifier, which can filter it out. In the DFLL, the DCO frequency is adjusted (in either increasing or decreasing fashion, depending on the initial conditions) until a comparator toggle is detected which indicates a successful lock due to V_{out} reaching V_{in}. Ideally, $\Delta V_{out} \approx 0$ and the DCO has a fine resolution ($\Delta F_{DCO}/\text{LSB} \to 0$), so $V_{out,0}$ will exactly equal V_{in} and the locking error will tend to 0. In reality, ΔV_{out} and the finite DCO resolution will cause the DFLL to lock to a range of voltages ΔV_{lock} whose corresponding frequencies ΔF_{lock} span above and below F_0. The voltage locking range is represented by

$$\Delta V_{lock} = V_{in,cmp} \times \frac{1+2\beta}{2\beta\,(1+\beta)} \approx \Delta V_{out}\Big|_{V_{out,1}=V_{in,cmp}} \tag{79}$$

where $\beta = C_L/C_S$ is the ratio of the capacitors used in the FVC. If the effects of temperature variation are ignored ($\Delta T = 0$), the (worst-case) frequency inaccuracy contribution from ΔF_{lock} can be expressed as

$$\frac{\Delta F_{lock}}{F_0} = \pm\frac{1}{1+2\beta} \tag{80}$$

Intuitively, (80) reflects that the frequency inaccuracy caused by the locking mechanism can be reduced by increasing the sizing of C_L relative to C_S which reduces the output voltage ripple of the FVC. Locking inaccuracies of a few hundred ppm are possible by using a large β around 1000, but this can impose a large chip area for C_L. A solution for the increased area of C_L is to use some quantity N of FVC in parallel with multi-phase operation ($\phi_1, \phi_2, ..., \phi_{2N}$) which allows the capacitor sizes of both C_L and C_S to be reduced while maintaining a similar ΔV_{out} and effective FVC transfer function [100]. Practically, the frequency variation due to locking inaccuracy is lower than (80) since the final proximity of F_{out} to F_0 depends on the dynamic probability of the comparator toggling as V_{out} moves closer toward V_{in} at each step of the locking process. An example is shown in Fig. 4.25(b) for a decreasing frequency search, where the probability of the DFLL detecting a lock is set by the percentage of time that V_{out} is above V_{in} during one cycle of the F_{in}. Assuming that $F_{dig} \ll F_{out}/N$ and that its phase is uncorrelated with F_{out}, the probability of finding a lock (i.e., the comparator samples V_{out} when it is above V_{in}) at a given frequency $F_{out} = f$ can be represented as a binomial experiment of n independent trials where n is the number of cycles of F_{dig}

$$P(\text{lock}) = 1 - \left(\frac{fC_L R_{ref}}{N} - \frac{2\beta C_L}{C_S(1+2\beta)}\right)^n \tag{81}$$

Then, if we consider a quantity Q of discrete and uniformly-distributed frequencies $f = \{f_1, f_2, ..., f_Q\}$ obtainable from the DCO over the interval ΔF_{lock} (i.e. $\Delta F_{DCO}/\text{LSB} = f_i - f_{i-1}$), the probability of locking to a frequency f_i can be calculated as

$$P(F_{out} = f_i) = P(\text{lock})\big|_{f_i} \prod_{k=1}^{i-1} \left(1 - P(\text{lock})\big|_{f_k}\right) \tag{82}$$

An example of the resulting probability distribution is shown in Fig. 4.25(c). For a DFLL with a double-sided locking approach, the joint probability is bimodal due to the assumed equal

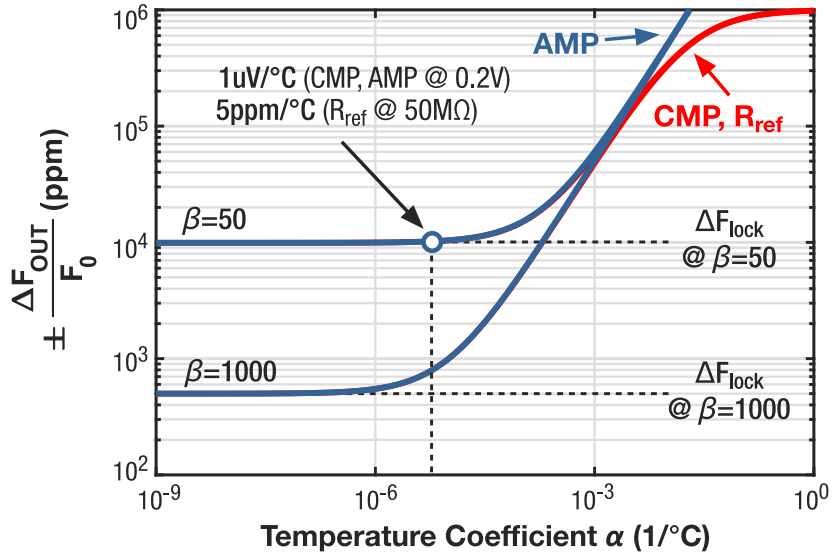

Fig. 4.26. Independent contributions of the amplifier, comparator, reference resistor, and locking circuitry to the total frequency inaccuracy of the DFLL, shown as a function of the TC of each component. A temperature range of 100℃ (ΔT=50) is used.

possibility of a lock occurring from either side of V_{in}. Increasing n or the DCO resolution will increase the likelihood that the DFLL will lock towards the boundaries of ΔV_{lock}, while increasing β towards infinity will eventually merge the two modes together to a single point at F_0 with probability of 1. To ensure an accurate DFLL lock within the limits of (80), the DCO resolution should be high enough that V_{out} can be be driven within ΔV_{lock}:

$$\frac{\Delta F_{\text{DCO}}}{\text{LSB}} \leq 2\left|\Delta F_{\text{lock}}\right| \tag{83}$$

This constraint can also be used to define the DCO accuracy requirements. Nonlinearity and non-monotonicity don't preclude a successful lock within ΔV_{lock}, but the differential nonlinearity (DNL) can cause V_{out} to jump completely across ΔV_{lock}(rather than settling within it) if its value $\left|DNL\right| \times (\Delta F_{\text{DCO}}/LSB)$ violates the condition in (83). Including the effects of temperature variation, the total steady-state output frequency inaccuracy can be expressed as a double-sided value

$$\frac{\Delta F_{\text{out}}}{F_0} = \pm\left(\frac{1+\alpha_{\text{amp}}\Delta T}{(1+2\beta)(1+\alpha_{\text{cmp}}\Delta T)(1+\alpha_{\text{R}}\Delta T)} + \left|\frac{\left(\alpha_{\text{cmp}}-\alpha_{\text{amp}}+\alpha_{\text{R}}\right)\Delta T+\alpha_{\text{cmp}}\alpha_{\text{R}}\Delta T^2}{(1+\alpha_{\text{cmp}}\Delta T)(1+\alpha_{\text{R}}\Delta T)}\right|\right) \tag{84}$$

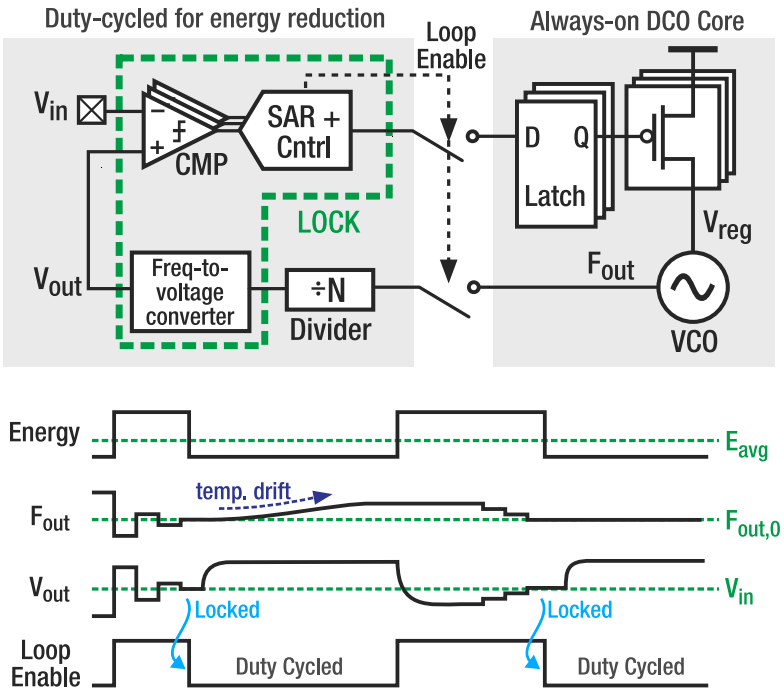

Fig. 4.27. Architecture and timing waveforms of the proposed duty-cycled DFLL.

Note that when $\Delta T = 0$, (84) becomes equal to (80). Fig. 4.26 shows how the frequency inaccuracy across ΔT=100 (converted to ppm by multiplying (84) by 1,000,000) is affected by the individual TCs of the amplifier, comparator, and reference resistor. The results for each component are obtained by sweeping its TC while holding the other TCs at 0. If the TC of any component can be decreased enough, the total frequency inaccuracy becomes limited by the locking inaccuracy in (80). When β=50, this corresponds to a TC of approximately 10^{-5}, which translates into 5ppm/℃ for a 50MΩ R_{ref}, or an offset drift of 1μV/℃ with a 0.2V V_{in} for both the amplifier and comparator. If the TCs are instead swept to infinity, the frequency inaccuracy will tend to infinity for α_{amp}, or 10^6 for α_{cmp} and α_R since they will eventually force F_{out} to 0Hz.

4.4.3 Duty-Cycling Concept

The proposed duty-cycled DFLL architecture, shown in Fig. 4.27, has two key mechanisms for reducing energy. First, the digital implementation improves the ability to operate robustly at low V_{dd} to take advantage of near/sub-threshold digital supply voltage rails in the 0.4-0.6-V range that are commonly used in IoT SoCs for low-energy digital circuits. Designing the DFLL

at V_{dd}=0.5V for compatibility with this regulated voltage domain will allow direct integration with low-energy digital circuits, reduce energy consumption, and eliminate the need for the DFLL to maintain stability over a wide range of supply voltages. Second, an integrated timing controller is used to duty-cycle the divider and locking circuitry (shown by the "loop enable" signal) for some duration t_{DC} once the frequency lock is complete. While duty-cycled, the DCO control word is frozen and the divider is disabled which breaks the feedback path and lets the DFLL run open-loop, which leaves it susceptible to temperature variations based on the intrinsic temperature stability of the DCO. To compensate for temperature drift, the DFLL is periodically reactivated to re-lock the output frequency before returning back to the duty-cycle state. The duty-cycled divider energy can be neglected since its leakage power is very low relative to the power of other FLL components. Within the locking circuitry, the DC power consumption from I_{ref} will decrease since the impedance of the FVC tends to infinity when $F_{in} = 0$, which effectively decreases m by 1. Further reduction could be achieved by power-gating the remaining circuitry in the FVC, such as the V-I converter. The factor of reduction in m during the duty-cycled state is modeled as a parameter $\gamma_{Iref} = m_{DC}/m_{nominal}$. A similar effect occurs in the digital circuits (comparators and control block) since the duty-cycled state reduces their switching activity, which effectively decreases C_{dig} by a factor γ_{dig}. When a duty-cycle ratio $r_{DC} = t_{on}/(t_{on} + t_{DC})$ is used (r_{DC}=1 is always active), the average energy per cycle of the DFLL can be expressed as

$$E_{cycle,avg}(N) = (r_{DC} 2C_{ff} + 5C_{inv})\, V_{dd}^2 + \frac{C_S R_{ref}}{N}\left(N I_{leak} V_{dd} + C_{dig} F_{dig} V_{dd}^2 (\gamma_{dig} + r_{DC} - \gamma_{dig} r_{DC}) + \frac{m V_{dd} V_{in}}{R_{ref}} (\gamma_{Iref} + r_{DC} - \gamma_{Iref} r_{DC}) \right) \quad (85)$$

Fig. 4.28(a) shows the energy consumption during active and duty-cycled operation of a DFLL with with V_{DD} = 0.5V, C_{INV} = 0.5fF, C_{FF} = 12fF, and an arbitrarily-chosen P_{lock} of 18nW. At low N, where the locking energy dominates, the amount of total energy reduction is limited by γ_{Iref} and γ_{dig}. However, if the energy of the locking circuitry is reduced by increasing N (or

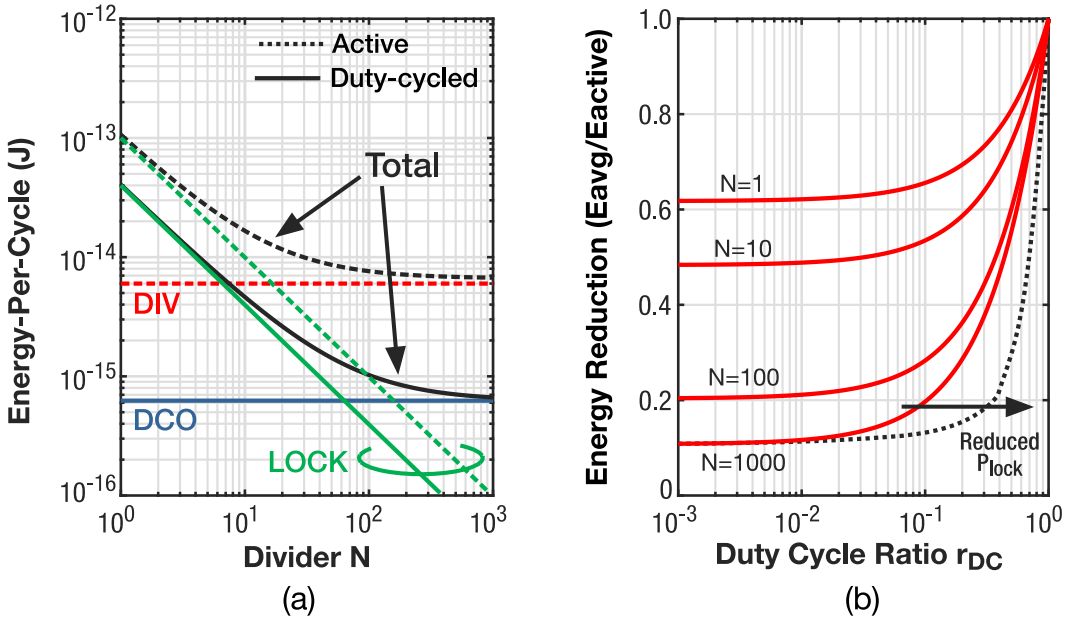

Fig. 4.28. DFLL energy consumption during active and duty-cycled modes (a). DCO energy is the same for both active and duty-cycled modes, and the divider energy is only present in the active mode. (b) shows total energy reduction ((85) divided by (74)) versus duty cycle ratio, shown for different divider values.

by reducing its power via reduced V_{dd}, C_{dig}, etc), then the total energy can reduced down to the power of the DCO. Note this is is in contrast with the analog FLL from Section 4.3 which is limited by the power of the divider. Fig. 4.28(b) shows that the average energy reduces with the duty-cycle ratio and will eventually plateau once the total energy is equal to the DCO energy. Note that the specific characteristics between energy reduction and r_{DC} is based on the relative strength of P_{lock} in Fig. 4.28(a). If the locking energy is lower to begin with, the the average energy will reach the same plateau at higher duty-cycle ratios (example shown by dashed line for N=1000).

4.4.4 Implementation

Fig. 4.29 shows the full schematic of the proposed duty-cycled DFLL design, which targets a 0.5V V_{dd} due to its popularity as a near/sub-threshold supply voltage rail. The DCO is implemented using a tunable PMOS array that adjusts the supply voltage V_{REG} of a 5-stage RO which is designed with minimum-sized devices for low energy (C_{INV}). The RO presents a very small load current, so the PMOS array uses long-channel thick-oxide devices. A skewed-sizing inverter restores the DCO output level to full swing. This level conversion

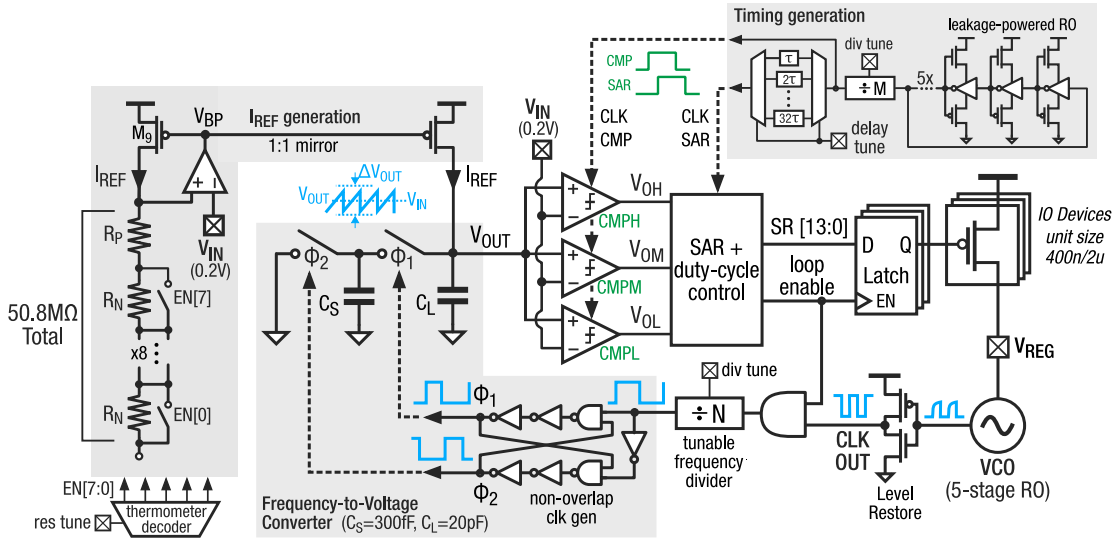

Fig. 4.29. Full schematic of the duty-cycled DFLL including I_{REF} generation circuit and always-on digital timing generation block.

approach does dissipate some short-circuit current, but also has less dynamic power than a more complex design with less short-circuit current. The level converter is switched at a high frequency, so the reduction in dynamic power is most preferable in this case. The FVC is implemented as shown in Fig. 4.4, and R_{ref} is compensated using the same approach as in the Analog FLL (see Section 4.3.1) by using two different resistors R_P and R_N to compensate for their total series temperature coefficient α_R. At the optimum tuning setting, the minimum TC of R_{ref} is approximately $\alpha_R = 1.7 \times 10^{-4}$/°C. C_L and C_S are 20pF and 300fF, respectively, yielding a $\beta \approx 66$. Further decreasing C_S could yield a smaller β, but can increase relative impacts of temperature-dependent parasitic capacitances [78]. An always-on timing generation circuit uses a 5-stage leakage-based ring oscillator (DLS inverters) and tunable frequency divider to generate F_{dig}, which is delayed to create separate clocks for the comparators (Clock (CLK) Comparator (CMP)) and digital control block (CLK SAR). The delay line simply ensures that the comparator outputs are valid at each clock cycle of the digital controller, and is designed using different numbers of the same leakage-based inverters in series. To achieve both a fast initial lock and smooth subsequent re-locking, the digital algorithm replaces the single comparator approach used in existing DFLL designs ([80, 82]) with a bank of 3 comparators that enables a hybrid binary/linear searching algorithm which will be discussed shortly. Fig. 4.30(a) and (b) show the schematics of the amplifier used in

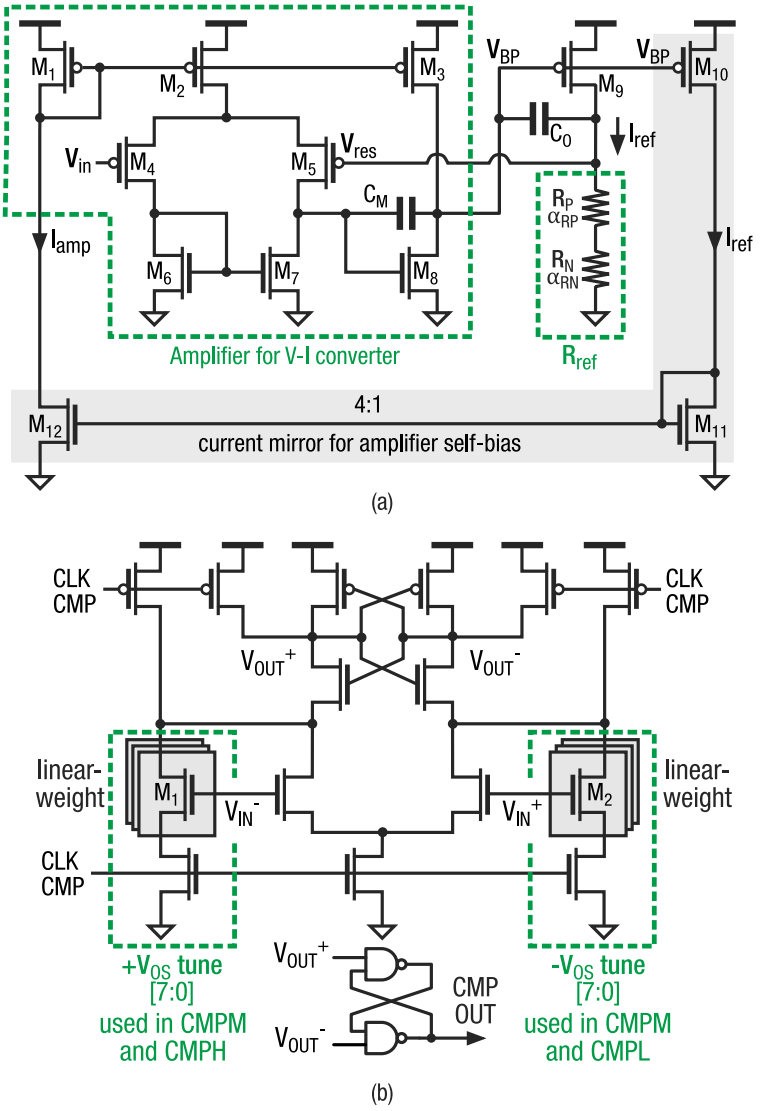

Fig. 4.30. Schematic of (a) the V-I converter with self-biased amplifier and (b) the clocked comparator based on a StrongArm latch with tunable offset.

the V-I converter and the comparator design, respectively. The amplifier uses a two-stage topology that is self-biased with a copy of the reference current I_{REF} that it generates. The offset voltage of the amplifier shifts by a few mV across the temperature range and equates to a TC of $\alpha_{amp} = 1.0 \times 10^{-4}$/°C. Note that it is ideal for this value to match α_R, since they cancel each other out in (84). The comparator is based on a StrongArm latch, where the offset voltage is tuned by adjusting the widths of the input pair which are implemented with 8-bit arrays of thermometer-coded minimum-sized devices that allos for a maximum tuning range of around 20mV. At baseline, the TC of this offset voltage is smaller than α_R and

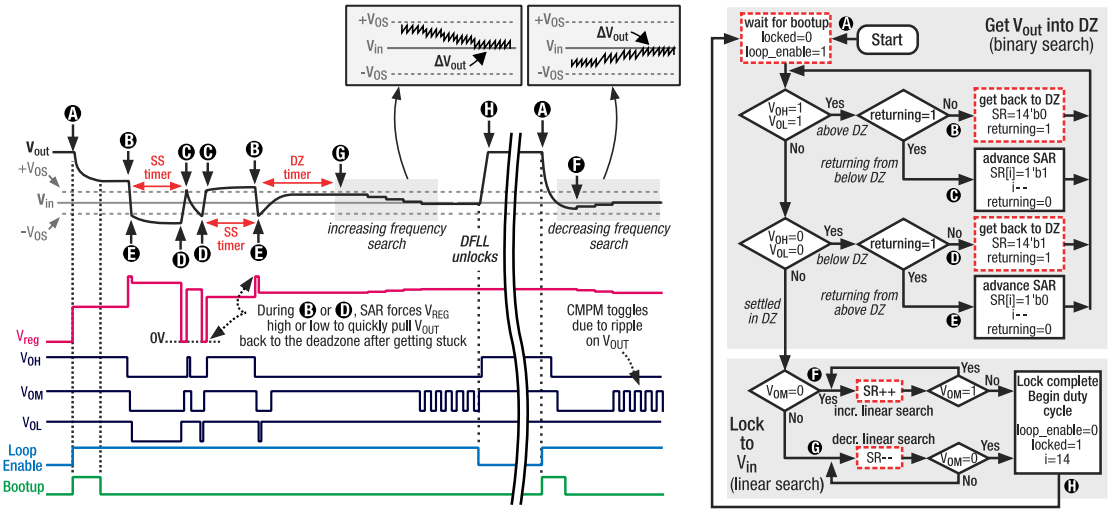

Fig. 4.31. Timing diagram of the DFLL booting up, locking, duty-cycling, and waking up to re-lock. Diagram of locking and duty-cycling algorithm shown on right, where processes with dashed borders indicate that they only proceed after waiting for a timer.

α_{amp}, roughly around $\alpha_{cmp} \approx 1 \times 10^{-5}$ such that its effect on (84) is negligible. Based on the individual component TC values, (84) predicts that across 100°C, the DFLL will obtain a total inaccuracy of 111ppm/°C. This value decreases to 34.7ppm/°C if β is set to 1000, which aligns well with measured results from the previously demonstrated analog FLL that does not suffer from a locking inaccuracy [DT15]. If all TCs are set to zero, the locking inaccuracy alone will lead to an overall TC of 81ppm/°C.

The digital locking algorithm uses three comparators (CMPH, CMPM, and CMPL) to achieve a fast frequency lock by establishing a voltage Deadzone (DZ) around V_{in}. The middle comparator CMPM is designed to have low offset to directly lock V_{in} with V_{out}, while CMPH and CMPL are tuned for positive and negative offset, respectively, to set the high and low bounds of the DZ. Upon bootup, a SAR binary search algorithm looks at the outputs of CMPH and CMPL and quickly adjusts the DCO control word (SR) until V_{out} is within the deadzone, and then a linear searching algorithm completes the locking process by adjusting the DCO one Least-Significant Bit (LSB) at a time until it detects a toggle on the output of CMPM, indicating that V_{out} has reached V_{in} (see diagram in Fig. 4.25). The frequency response of the FVC can be studied by treating its input as V_{in} of the V-I converter, while

assuming F_{in} is constant:

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{1/(R_{ref}F_{in}C_S)}{1+\frac{C_L}{C_S F_{in}}s} \tag{86}$$

Intuitively, (86) has a DC gain consistent with the result from (45), and single pole at $-C_S F_{in}/C_L$ with a notable dependence on F_{in}. Changes in F_{in} during the binary searching algorithm (while V_{in} is constant) are analogous to a step input on V_{in} with a constant F_{in} which elicits a characteristic RC response. Therefore, to ensure the FVC has time to settle after each bit trial, the algorithm's speed (F_{dig}) should be in the pass-band of the FVC. Since the FVC cutoff frequency has a dynamic dependence on F_{in}, the locking algorithm uses several configurable timers to effectively adjust its response time based on its current stage which is a proxy for the current range of F_{in}. A timing diagram and flow chart of the locking, re-locking, and duty-cycling processes are detailed in Fig. 4.31. When the DFLL first boots up, it performs the first bit trial and waits on a bootup timer before using CMPH and CMPL to make the first decision (step A). After each bit trial of the binary search, a Steady-State (SS) timer begins so that the FVC can settle. If V_{out} is outside the DZ ($V_{OH} = 1 \& V_{OL} = 1$ or $V_{OH} = 0 \& V_{OL} = 0$) when the SS timer expires, this indicates that V_{out} is stuck outside the DZ and the algorithm responds by setting the DCO control to its maximum or minimum value to quickly return V_{out} to the the DZ (steps B and D) before moving to the next bit trial (steps C and E). If V_{out} enters the DZ at any point, a deadzone timer begins which will trigger the linear searching algorithm if V_{out} has not left the DZ when the DZ timer expires. Depending on whether V_{out} is above or below V_{in} when the linear searching algorithm begins (steps F and G), it will begin increasing or decreasing the DCO frequency, respectively, until it detects a toggle on the CMPM output. For each step in this phase, a linear-searching timer limits the number of times V_{out} is sampled by CMPM (represented by n in (81)) before the algorithm increments or decrements the DCO frequency. Once the lock is complete, the DCO control word is frozen and the feedback loop is disabled so that V_{out} is unregulated and charges to V_{dd}, and a new duty-cycle timer starts to limit the duration of the duty cycle interval to $t_{DC} = n_{DC} T_{dig}$, where n_{DC} is the counter threshold of the duty-cycle timer.

While duty-cycled, the temperature sensitivity and noise performance of the DFLL are based on the free-running DCO whose frequency increases exponentially with temperature

but can be linearized at F_0 as

$$f_{\text{DCO}} = F_0 + \alpha_{\text{DCO}}\Delta T \tag{87}$$

where α_{DCO} is given in Hz/°C and is simulated to be around 15,000. Depending on the duty-cycle ratio and the rate of ambient temperature fluctuations, the open-loop DCO temperature sensitivity can contribute additional inaccuracy to the DFLL which will be discussed shortly. When the DFLL wakes up to re-lock, the bootup timer allows the FVC to settle, whose output will have deviated from V_{IN} by some amount $\Delta V_{\text{out,0}}$ based on how far the temperature has drifted since the last frequency lock:

$$\Delta V_{\text{out,0}} = \alpha_{\text{DCO}}\Delta T \left.\frac{\partial V_{\text{out}}}{\partial F_{\text{out}}}\right|_{\text{F}_{\text{out}}=\text{F}_0} \tag{88}$$

$$= \frac{-\alpha_{\text{DCO}} C_{\text{S}} R_0 V_{\text{in}} \Delta T(1+\alpha_{\text{amp}}\Delta T)}{N(1+\alpha_{\text{R}}\Delta T)} \approx 5\text{mV/C} \tag{89}$$

This value plays a role in determining the amount of temperature drift that can be tolerated such that V_{out} will still be in the DZ when the DFLL wakes up to re-lock. Once the FVC has settled, the controller samples V_{out} from CMPM and then waits the duration of the linear-searching timer for a CMPM toggle which would indicate that V_{out} is still locked and no DCO adjustments are necessary. If a toggle is not detected, the linear searching process begins and can complete the re-locking process regardless of whether V_{out} is in the deadzone or not. This approach allows the re-locking process to occur in the background without causing the large jumps or transitions in F_{out} as in the binary algorithm. If V_{out} is outside the DZ after bootup, the controller can optionally begin the binary searching algorithm instead.

The FVC output can be modeled in continuous time with the simple first order difference equation

$$C_{\text{L}}\frac{dV_{\text{out}}(t)}{dt} + \frac{V_{\text{out}}(t)}{Z_{\text{FVC}}} = \frac{V_{\text{in}}}{R_{\text{ref}}} \tag{90}$$

and the settling time t_{settle} of the FVC when waking up from the duty-cycled state can be solved with (90) knowing that the initial value of V_{out} is V_{dd} and that V_{out} can be considered

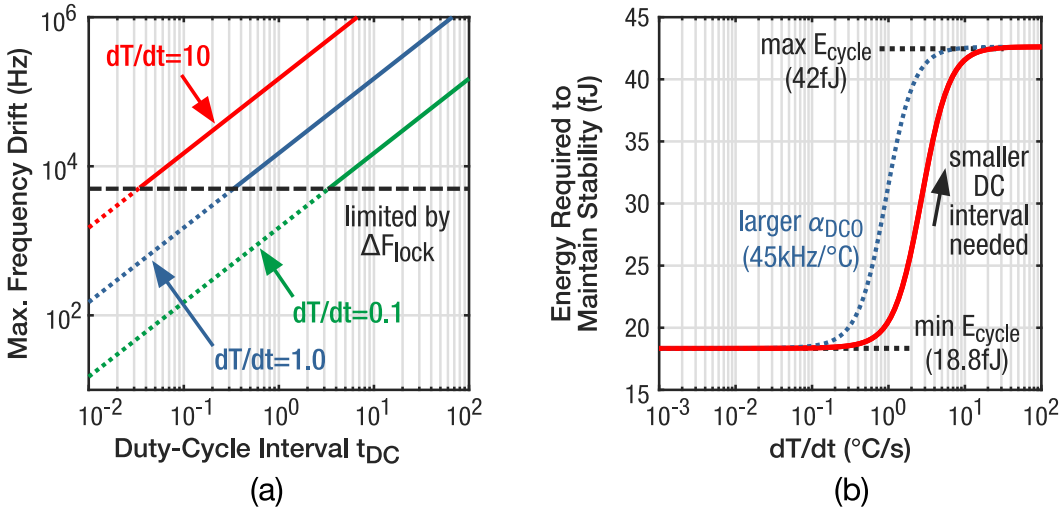

Fig. 4.32. Modeled maximum open-loop frequency drift versus duty-cycle interval shown for temperature fluctuations of 0.1°C/s, 1°C/s, and 10°C/s (a), and average energy per cycle of the DFLL required (based on duty-cycle rate) to limit open-loop frequency drift to within ΔF_{lock} versus the rate of ambient temperature drift (b).

settled when it reaches the deadzone:

$$t_{settle} = \frac{C_L}{C_S F_{in}} \ln\left(\frac{C_S F_{in} R_{ref} V_{dd} - V_{in}}{C_S F_{in} R_{ref} (V_{in} + V_{os}) - V_{in}}\right) \tag{91}$$

where V_{os} is the CMPH offset voltage and F_{in} will be f_{DCO}/N which accounts for the temperature drift of the DCO frequency while duty-cycled. The overall temperature dependence of t_{settle} is low, and its value at N=10 in the proposed implementation is around 5ms. Once the FVC has settled, the linear search must complete the lock. The total re-locking time can then be approximated by

$$t_{relock} = t_{settle} + (n_{linear} T_{dig}) \frac{\alpha_{DCO} \Delta T}{\Delta F_{DCO}/\text{LSB}} \tag{92}$$

where n_{linear} is the counter threshold used for the linear searching timer that corresponds to n used in (81). The temperature dependence of (92) is essentially linear and limited at $\Delta T = 0$ by t_{settle}. The time required by the linear searching algorithm to re-lock can be minimized by running as fast as possible within the limit of (91) due to the FVC settling time. Taking these factors together, typical locking and re-locking times on the order of several milliseconds can be expected.

4.4.5 Transient Frequency Inaccuracy

Short-term transient inaccuracy caused by frequency drift within a single duty-cycle interval can be evaluated with (87) by replacing ΔT with $\Delta T = dT/dt \times t_{DC}$, where dT/dt is the instantaneous rate of temperature change expressed in ℃/s and t_{DC} is the length of time spent in the duty-cycled state. Fig. 4.32(a) shows the resulting frequency drift versus duty-cycle interval for multiple rates of temperature change. In some cases, the frequency drift of the DCO will be lower than ΔF_{lock}, which will still limit the total frequency inaccuracy. To prevent transient frequency inaccuracy from exceeding ΔF_{lock}, the necessary t_{DC} can be computed for any dT/dt and combined with (92) to obtain the effective duty-cycle rate $r_{DC} = t_{relock}/(t_{relock} + t_{DC})$. Using the resulting r_{DC} with (85) and (74) yields the average energy per cycle of the DFLL required to maintain the intrinsic steady-state frequency inaccuracy by limiting the open-loop frequency drift. This result is shown in Fig. 4.32(b). For this DFLL design with α_{DCO}=15kHz/℃, low duty-cycle rates that minimize the average energy can be used until ambient temperature fluctuations exceed 1℃/s, at which point faster duty-cycle intervals are required that will increase energy toward the fully-active value. Variations that increase α_{DCO} will require shorter duty-cycle rates and therefore higher energy at lower rates of temperature change, as shown by the dashed line.

4.4.6 Measurement Results

The proposed duty-cycled DFLL was fabricated in a low-power 65nm process, and Fig. 4.33 shows an annotated chip micrograph. The design occupies an area of 0.134mm^2, of which roughly 60% is consumed by R_{ref}. Fig. 4.34 shows the measured power and energy consumption of the FLL versus its output frequency range while a 99% duty cycle is used. Since this design uses the same fundamental values that set the frequency (R_{ref}, C_s) as the previously demonstrated analog FLL, the output frequency range is similar with a base frequency of 55kHz (N=1). At low N, the locking power dominates, so the total power remains nearly constant while energy changes inversely with increasing frequency. However, the DCO (and level restorer) becomes increasingly dominant as N is increased, causing a proportional increase in power that causes energy to plateau, where it reaches 18.8fJ/cycle at N=10, when

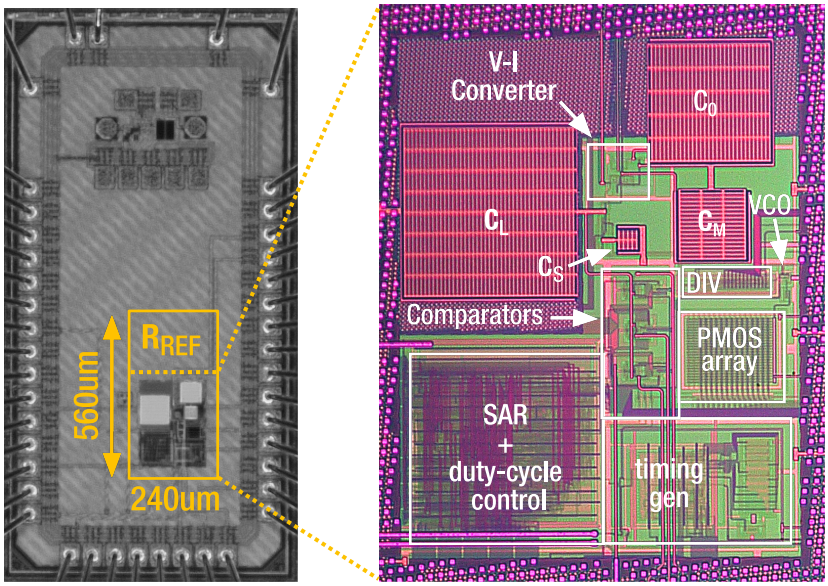

Fig. 4.33. Annotated chip micrograph of the proposed DFLL in 65-nm CMOS.

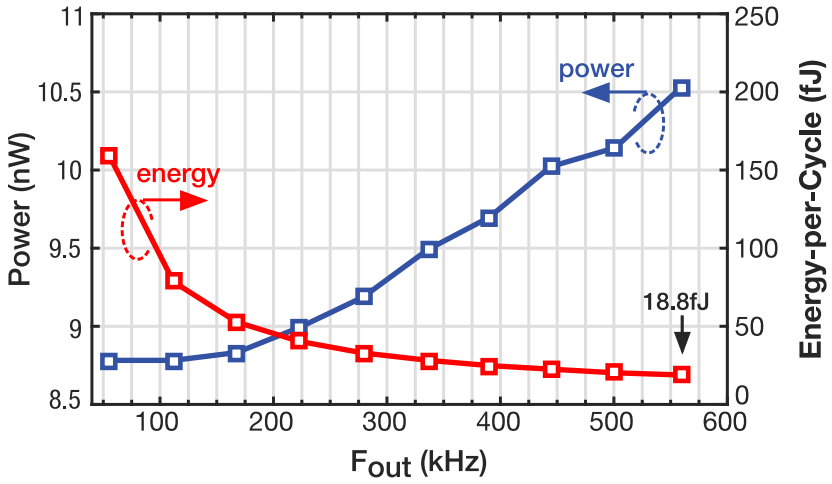

Fig. 4.34. Measured and power and energy consumption versus output frequency F_{out} measured at 20°C.

F_{OUT}=560kHz. Without duty-cycling, the total active energy at this point is approximately 42fJ/cycle. This can be compared to the performance to the analog FLL which consumed 34nW at 580kHz , yielding 58fJ/cycle. Note that the analog FLL could reach higher F_{out} values (higher N) due to a combination of its higher V_{dd} and its VCO design allowing a different range than the DCO used for this design. The DFLL was tested across a temperature range of 0°C to 100°C in a Tenney Jr. environmental temperature testing chamber, and the output frequency was recorded at 10°C increments after allowing the temperature to fully settle (no

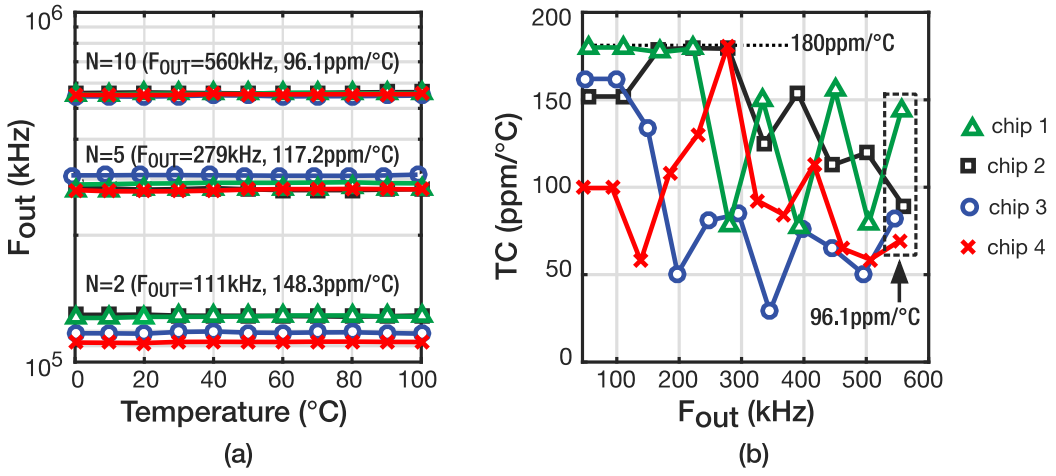

Fig. 4.35. Measured output frequency F_{out} versus temperature, shown for divider values N=2, N=5, and N=10 on four dies (a) from 0°C to 100°C. TCs of the four measured dies are shown across the full range of output frequencies (N=1 to N=10) (b).

transient frequency inaccuracy). This process was repeated for divider N values ranging from 1 to 10 on four separate dies. Fig. 4.35(a) shows the measured frequency outputs of the four dies across the temperature range for divider values of N=2, N=5, and N=10. The resulting steady-state TC for all possible divider values (N=1 to N=10) is shown in Fig. 4.35(b) and demonstrate full functionality across temperature from 55kHz to 560kHz, with a worst-case TC of 180ppm/°C between all dies and divider values. This limit was separately observed in multiple dies and is likely due to a worst-case combination of locking inaccuracy and temperature coefficients in the amplifier and R_{ref}. The average TC among all output frequencies is closer to the predicted value of 111ppm/°C , and is measured as 96.1ppm/°C when F_{out}=560kHz.

Fig. 4.36 shows the transient frequency inaccuracy of the DFLL when subjected to fluctuations in temperature. A heating element is attached to the chip package and cycled on/off in 10-minute periods which induce a +/- 20°C fluctuation with a peak measured dT/dt of 0.5°C/s. The DCO is immediately susceptible to the increasing temperature as shown by the open-loop response which increases up to 1.1MHz. The transient closed-loop response during the same temperature change is shown for duty-cycle intervals of t_{DC}=0.1s, t_{DC}=1s, and t_{DC}=8s, which correspond with duty-cycle rates of r_{DC}=0.36, r_{DC}=0.06, and r_{DC}=0.01, respectively, based on the measured t_{relock}. Shorter duty-cycle intervals reduce the amount

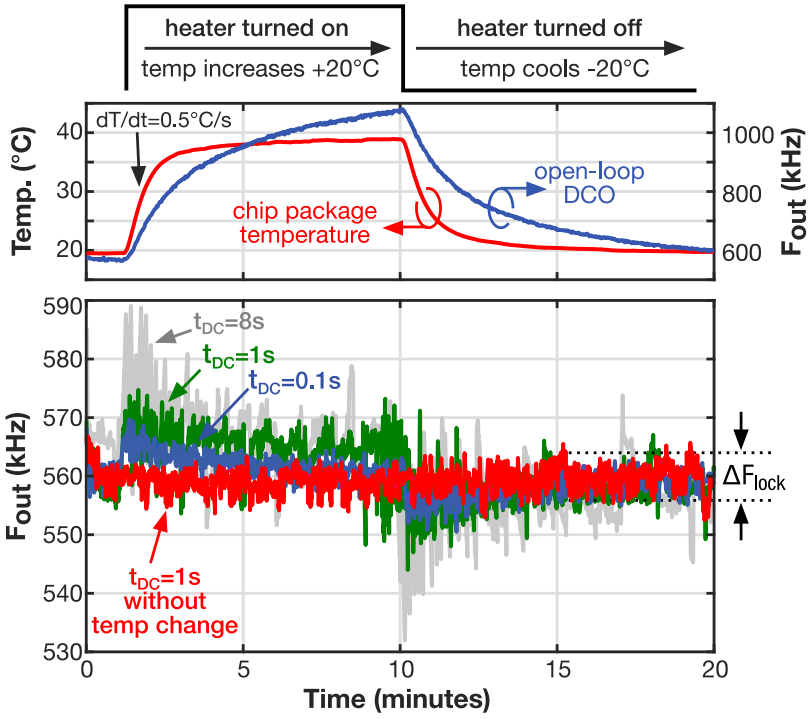

Fig. 4.36. Measured transient operation of the DFLL during a positive and negative 20℃ step in temperature, shown for duty cycle rates of t_{DC}=0.1s (r_{DC}=0.36), t_{DC}=1s (r_{DC}=0.06), and t_{DC}=8s (r_{DC}=0.01). For reference, the open-loop DCO response (with temp. change) and baseline closed-loop operation (t_{DC}=1s without temp. change) are shown.

of frequency drift within each duty-cycle interval that results in a lower transient inaccuracy that converges toward the steady-state inaccuracy defined by ΔF_{lock}. When the heater turns off and the chip package starts to cool, dT/dt becomes negative and the DCO frequency decreases and swings to the other side of the deadzone, so the DFLL compensates during the re-locking process by increasing the DCO frequency back to the deadzone. The discrepancy in F_{OUT} during the heating and cooling phases highlights the inaccuracy caused by ΔF_{lock}.

Frequency stability in the time domain is shown by the Allan deviation plot in Fig. 4.37, which was captured at room temperature over a 27-hour long measurement period for duty-cycle intervals of t_{DC}=1s and t_{DC}=14s. For reference, the open-loop DCO is also shown. Correlated DCO noise limits stability at less than 1s regardless of the duty-cycle interval, while longer-term stability up to 12 hours can be held to several hundred ppm by duty-cycling to filter out temperature variation. Fig. 4.38(a) shows the measured supply sensitivity of the DFLL across a ±50mV range from the nominal value. The total variation across this range (1.1%)

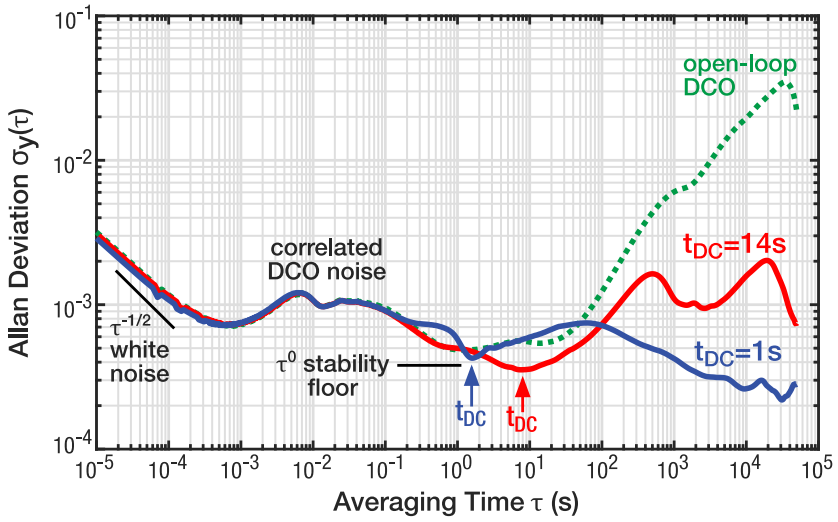

Fig. 4.37. Allan deviation measurement of the DFLL at F_{out}=560kHz (N=10) in an indoor environment at room temperature (20℃), shown for open-loop operation as well as duty-cycle intervals of 1s and 14s.

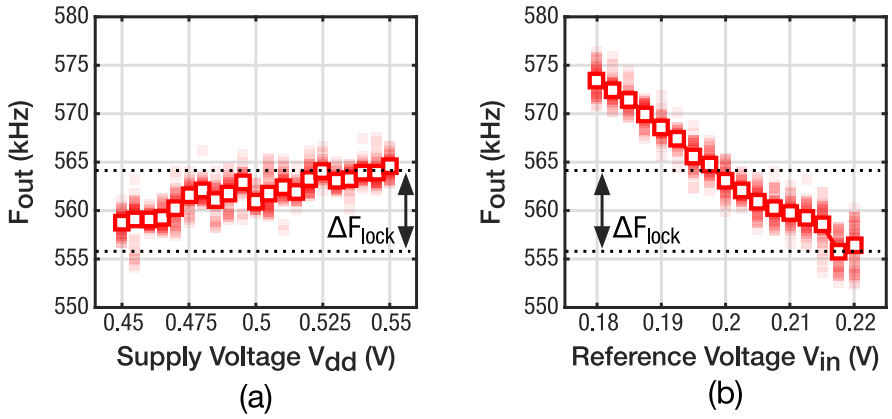

Fig. 4.38. Measured supply voltage sensitivity for ±50mV of deviation from the nominal supply voltage (a), and reference voltage sensitivity for ±20mV of deviation from the nominal reference voltage. Both measurements taken at 20℃.

equates to a supply sensitivity of 11.01%/V. The models used in this dissertation predict that the FLL architecture is theoretically robust to the input voltage V_{in}, and measurements from the Analog FLL in Section 4.3 showed that ±10mV have little effect on F_{out} the temperature range. We repeat this assessment at room temperature across a V_{in} variation of ±20mV for the DFLL by booting it up and letting it run for 50 consecutive duty-cycle intervals, measuring F_{OUT} at the end of each interval. The results, shown in Fig 4.38(b), demonstrate a small linear dependence of F_{out} on V_{in}. Small variations within a 10mV range have a negligible impact compared to the locking inaccuracy, which is modeled in (80) and demonstrated in

Fig. 4.36. For V_{in} to remain in this range, a temperature stability of less than 475ppm/℃ is required, which can easily be achieved by a simple voltage reference design [101].

	This Work	[82]	[80]	[83]	[DT15]	[79]	[78]
Architecture	DC-DFLL	DFLL	DFLL	DFLL	FLL	FLL	FLL
Technology (nm)	65	40	40	180	65	180	180
Area (mm^2)	0.134	0.07	0.07	1.59	0.098	0.16	0.26
Supply Voltage (V)	0.45–0.55	1.0–1.4	0.65–0.8	1.7–2.0	0.6–0.8	1.0–1.8	1.2–1.8
Frequency (kHz)	560	428	417	7000	1016	32.7	70.4
Temp. Range (°C)	0–100	-40–80	-20–80	-45–85	-20–60	-20–100	-40–80
Trim Method	Res.	Time Domain	Res.	Res.	Res.	Res. + Cap.	Res.
Num. Trim Points	1-point	13-point	1-point	2-point	2-point	2-point	2-point
Temp. Stability (ppm/°C)	96.1	8	135.3	2.5	20.3	13.2	34.3
Num. Sampled Dies	4	3	3	12	1	4	5
Power (nW)	10.5	380	181	775000	45.3	35.4	110
Energy per Cycle (fJ)	18.8	900	430	110000	44.6	1080	1560

Table 11. Performance summary of digital FLL design and comparison to state-of-the-art. Note that comparison is limited to works that were available at time of publication (12/28/2020), and more recent works are not shown here. Reference publication for this work is [DT2].

4.4.7 Conclusion

This section presented a Digital Frequency-Locked Loop (DFLL) architecture that allows the feedback loop to be reliably broken so that energy-limiting components can be disabled (duty-cycled) for power and energy reduction and then reactivated periodically to compensate for ambient temperature drifts. Theoretical derivations of the energy efficiency predict the energy savings obtained from duty-cycling, and the steady-state and transient frequency stability of the design is also analyzed. The DFLL design is fabricated in 65-nm CMOS, and measurements results show that this design achieved the highest energy efficiency amongst all temperature-compensated oscillators (RXOs and FLLs published to date to the best of current knowledge. A reasonably high stead-state TC is achieved, and experimental results also confirm the expectation that spending longer amounts of time in the low-power duty-cycled state reduce the frequency stability during transient temperature fluctuations.

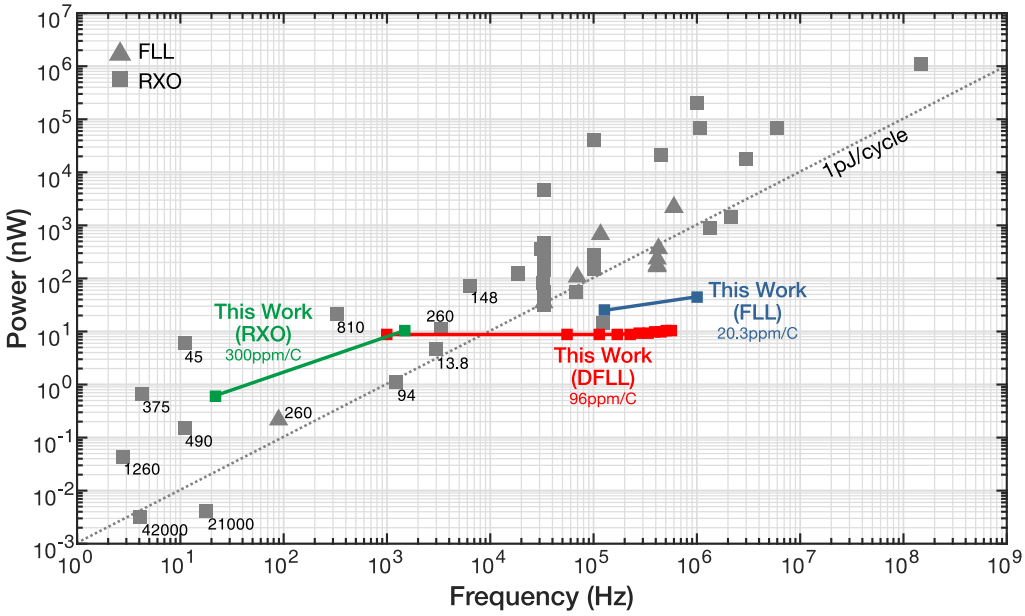

Fig. 4.39. Power versus frequency design space for on-chip clocks. Triangle markers show FLL designs, square markers show RXO designs. For very low-power designs, labels are printed to indicate the TC of the design.

4.5 Design Space Summary

Fig. 4.39 shows the current design space for on-chip clocks including the works presented in this section of this dissertation. As expected, it costs more power to operate at higher frequencies, and most designs sit near a 1pJ/cycle energy efficiency limit regardless of their operating frequency. Very low-power operation in the pW-range is possible for Hz-range frequencies, and RXOs are the dominant architecture in this area, but this performance typically comes at the cost of temperature stability which leads to TCs in the hundreds or thousands of ppm/°C. For frequencies above a few kHz, the DFLL from Section 4.4 is the most energy-efficient choice with increasing prominence at higher frequencies. The analog FLL becomes the preferred choice once the desired frequency increases beyond the maximum F_{out} limit of the DFLL which is supply-voltage limited. At Hz-range frequencies, the presented RXO is not the most energy efficient (mostly due to the 1V supply voltage), but it does achieve a reasonable TC compared with other works in that range. If the focus is instead shifted to temperature stability, a Pareto-optimal performance tradeoff can be shown between temperature stability and energy efficiency. This is shown by Fig. 4.40, which essentially visualizes the energy cost to obtain a certain TC. Intuitively, it costs more energy (more power

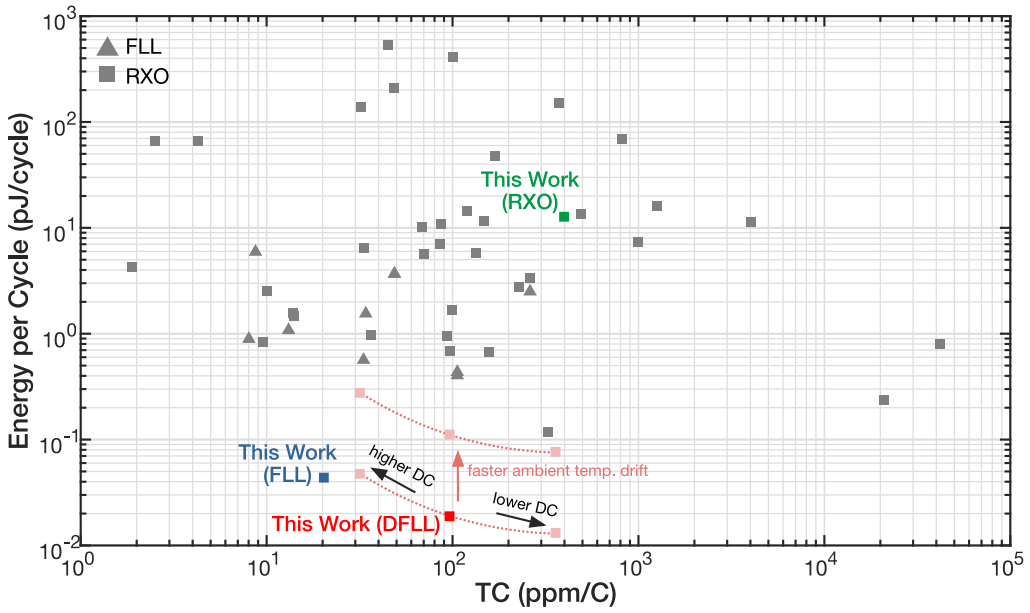

Fig. 4.40. Measured supply voltage sensitivity for ±50mV of deviation from the nominal supply voltage (a), and reference voltage sensitivity for ±20mV of deviation from the nominal reference voltage. Both measurements taken at 20°C.

relative to whatever the frequency happens to be) to achieve higher temperature stability, with the cost generally exceeding 1pJ/cycle to achieve a stability better than 10ppm/°C. Both of the FLL designs presented in this dissertation require the least energy cost for their respective TCs. Relative to existing works with similar temperature stabilities, these designs improve energy efficiency by at least 1 order of magnitude. An interesting note to make here is for the ability of the DFLL to dynamically trade off between energy and temperature stability. As shown in Fig. 4.32 and Fig. 4.28, the DFLL reduces energy by spending longer amount of time duty-cycled, but this increases its susceptibility to ambient temperature drifts. Depending on the worst-case rate of ambient temperature change, it may cost the DFLL more energy via a higher duty-cycle rate to maintain the same level of temperature stability.

5 Conclusion

The IoT is a powerful technological revolution that has solved many problems but also created many new ones as it has grown. Currently, there are several existing challenges and opportunities for the IoT that together give rise to the motivation for this dissertation. First, the scale and impact of the IoT are limited by the lifetime or longevity of the hardware sensor nodes on which it is built. Having nodes easily operate for longer amounts of time will increase the scale and application possibilities of the IoT. Second, nodes can last for indefinite amounts of time (to address the first issue) if they can harvest ambient energy at a rate that matches or exceeds the power consumption of their circuits. Third, there is growing support for energy harvesting from new sources (to better address the second issue), however the amount of power that can be harvested from these sources is quite low. Lastly, the power consumption of circuits in IoT nodes are limited by some major obstacles that are dependent on the classic architectures and techniques upon which they are built. In turn, there is a need to further reduce the power consumption of circuits (this has always been a relevant research problem) in IoT nodes to both increase performance under existing energy harvesting conditions and more importantly enable indefinite (perpetual self-powered) operation at low power from newer untapped energy harvesting sources. This technical advancement will then grow the scale and application range of the IoT. The contributions of this dissertation address the above problem statement by proposing, analyzing, modeling, and experimentally validating several fundamental design techniques to reduce power consumption in a variety of components that are essential to modern IoT nodes.

5.1 Summary of Results and Contributions

5.1.1 Digital Logic

- The operation of Dynamic Leakage Suppression (DLS) logic is analyzed, and several requirements are defined to ensure successful gate operation, including a new definition for switching margins (SMs) which are a new phenomenon that occur due to the input-

out hysteresis and threaten the ability of a gate to be reliably switched.

- Models for several fundamental performance metrics (input thresholds, gate delay, and leakage power) are derived and compared with simulated data.
- It is shown that a minimum leakage power point exists for DLS logic based on the relative significance of gate and subthreshold leakages of individual transistors. A derivation is created for the optimum supply voltage to minimize leakage power consumption in DLS logic.
- A test chip is fabricated in 65-nm CMOS with several unique DLS arrangements, or "flavors", based on design insights from the previously-derived models. These flavors make explicit tradeoffs between leakage, delay, and reliability. Measurement results both confirm expectations based on theoretical models and demonstrate proof that DLS achieves less leakage than HVT static-CMOS logic by nearly one order of magnitude .
- Measurements for within-die and die-to-die robustness provide further insight into the reliability of DLS logic.
- A new Scalable Dynamic Leakage Suppression (SDLS) logic family is proposed to increase performance flexibility beyond the limits of traditional DLS logic.
- Supporting hardware is designed to operate with the SDLS logic style that enable on-the-fly performance scaling via dynamic voltage adjustment and adaptive clocking to track the critical path delay.
- A RISC-V processor is implemented with an SDLS standard cell library, and a microprocessor test chip is designed that includes the supporting performance-scaling hardware that is accessible to the processor via a memory-mapped register. Jacob Breiholz led the RTL design and digital implementation for this chip, and Sumanth Kamineni and Ningxi Liu built the SRAM that was integrated on-chip.
- The SDLS microprocessor was fabricated in 65-nm CMOS and was experimentally demonstrated during on-the-fly performance scaling that dynamically tunes power and frequency while computing the Fibonacci sequence [DT17]. Ultimately, the SDLS

RISC-V core design achieves a minimum power of 840pW and can scale its frequency up to 41.5kHz.

5.1.2 Memory

- A new Static Random-Access Memory (SRAM) bitcell is proposed that leverages DLS inverters to form the cross-coupled inverter pair to reduce leakage power.
- An analysis of the operation of the DLS bitcell is performed, and it is shown that the low intrinsic drive strength of the DLS bitcell in combination with traditional peripheral circuits leaves it susceptible to data hold errors and read errors.
- A Word Line (WL) overdrive technique is proposed to reduce the leakage of the bitcell access transistors to prevent data hold errors. To create the overdrive voltage, a DLS-based level converter circuit is designed that boosts the WL select signal to a voltage above the bitcell V_{dd} [DT1]. The level converter consumes very little power so as to not mitigate the power savings of the DLS bitcell and is additionally very compact so as to not significantly increase the SRAM area.
- A 3-T read port is added to the bitcell to enable safe read access without corrupting the data. This contribution is made by Shourya Gupta.
- A full 16kb SRAM macro is designed for the DLS bitcell and fabricated in 65-nm CMOS [DT7]. Shourya Gupta performed the macro implementation, including layouts. Measurement results for the DLS SRAM chip show that the bitcell achieves a leakage of 614aW, and the full macro consumes less than 200pW and is operable from 0°C to 60°C.

5.1.3 Clocking Circuits

- For low-frequency clock generation, it is shown that the Relaxation Oscillator (RXO) is an energy-efficient architecture that is limited by static power of reference currents, and that increasing the value of the reference resistor R_{ref} to reduce power when Hz-range frequencies are desired would require a large amount of silicon area.

- It is shown that the effective resistance of a transistor's gate oxide, indirectly created by gate leakage current, can be used as an area-efficient replacement for the traditional reference resistor R_{ref}. The I-V characteristics and temperature dependence of this effective resistance are analyzed and modeled, and a voltage-to-current converter circuit is shown that can utilize this effective resistance to generate a constant reference current.

- It is shown that traditional methods for on-chip tuning of a reference resistors will fail if used with the proposed gate leakage-based resistance. A new circuit design is proposed that enables high-precision tuning of the effective resistance in order to trim its total value and to compensate its temperature dependence.

- A RXO is designed that uses the developed gate leakage-based reference current source and fabricated in 65-nm CMOS [DT13]. Measurement results show that the design can achieve sub-nW power consumption with sub-Hz frequency tuning and passive compensation to around 300ppm/°C from -40°C to 110°C by combining the gate leakage of different types of transistors. With an on-chip temperature sensor [DT20], dynamic temperature compensation can be performed to maintain 41ppm°C.

- An analog Frequency-Locked Loop (FLL) architecture is analyzed for kHz-MHz frequencies, and its energy consumption and temperature stability are modeled. An energy-efficient FLL design is fabricated in 65-nm CMOS and was operational from approximately 65kHz to 1MHz across a temperature range of -20°C to 60°C. The design achieves a peak temperature stability of 20.3ppm/°C, and a power consumption of 45.3nW at 1MHz.

- A duty-cycled Digital Frequency-Locked Loop (DFLL) architecture is proposed to reduce power and energy beyond the limits of traditional analog FLLs but also sacrifice on temperature stability. The energy efficiency and temperature stability are modeled, and a DFLL design is fabricated in 65-nm CMOS and was operational from approximately 55kHz to 560kHz across a temperature range of 0°C to 100°C. The design achieves a peak temperature stability of 96.1ppm/°C, and a power consumption of 10.5nW at

560kHz.

5.2 Future Work

Despite the fact that Dynamic Leakage Suppression (DLS) logic was invented over a decade ago, there has been very little foundational technical understanding of its operation prior to this dissertation. Previous work has shown successful experimental results and very brief analysis of operation, but it has remained in a relatively poorly understood state. The modeling and analysis provided by this dissertation offer a technical foundation to further explore DLS logic, and there is much more room to investigate. With a more complete theoretical understanding of DLS logic, it would now be useful to investigate how technology selection could affect its performance. The results here suggest that devices with low gate leakage and relatively higher subthreshold leakage provide the most reliable operation and the greatest reduction in power over traditional static-CMOS logic. It would also be interesting to explore whether devices can be optimized at the technology level, from the bottom-up, (doping, geometry, etc.) to best meet the needs of DLS logic. There is also room to explore the more fine details of device selection, sizing, and physical placement (with respect to layout-dependent effects) to improve reliability and yield. Based on the results from this work, it seems that greater reliability is needed from DLS logic before it could be successfully commercialized.

In the area of clocking circuits, the two FLL architectures presented in this dissertation stand on their own in terms of energy efficiency with respect to nearly all of the existing works. In other words, there is a very distinct performance area that is very clearly reachable but relatively unexplored. I expect that in the coming years, more designs will recognize this area of performance and attempt to reach it, and I believe that there are several possible adaptations of the techniques from this dissertation that can be leveraged in other designs and architectures to reach similar performance levels. For example, the only thing that separates a clock from a temperature sensor is the temperature coefficient of a resistor. As shown throughout Section 4, the output frequency of a clock is usually based on an RC time constant, where it is desired for R to have no temperature dependence so that the clock can maintain a steady frequency across temperature. Instead, if R can be designed with a robust and predictable temperature dependence, then the output clock frequency will similarly

change across temperature in a robust and predictable way, and if a stable frequency is available from a XO, then the output of the clock can simply be quantized to find a digital readout that reflects changes in temperature. This is a common approach used by many existing temperature sensors. Therefore, the ultra-energy efficient frequency generation approaches from this dissertation could be easily leveraged to create temperature sensors with similar levels of energy efficiency. The DFLL design, particularly the duty-cycled aspect of its operation, seems to be the key strategy for achieving high energy efficiency in the future in other analog components. There are a few improvements that could be made to the design presented here, one of which would be to add additional power gating to many of the sub-blocks while the design is duty-cycled. Specifically, the DFLL here leaves all of the analog components (I_{ref} generator, FVC, etc.) turned on even during the duty-cycled state, and they could be power-gated to further reduce power. Two key limitations of the frequency inaccuracy are the steady-state and transient errors. In the steady-state, the frequency-locking error ΔF_{lock} limits the maximum stability, so approaches to improve the frequency-locking resolution (without dramatically increasing the power consumption, chip area, etc.) are needed. Additionally, the transient frequency inaccuracy is a fundamental tradeoff of the design, where you are voluntarily allowing susceptibility to ambient temperature drifts while the loop is unlocked. It would be interesting to see if the design could be left normally unlocked with only a temperature “watchdog” (e.g., a very low power and relatively inaccurate temperature sensor) detecting if temperature has drifted enough to close and re-lock the loop before returning to open-loop operation.

Appendices

This page intentionally left blank.

A Published Works

[DT1] **D.S.. Truesdell** and B. Calhoun, "A single-supply 6-transistor voltage level converter design reaching 8.18-fJ/transition at 0.3-1.2-V range or 44-fW leakage at 0.8-2.5-V range," *IEEE Solid-State Circuits Letters (SSCL)*, vol. 3, pp. 502–505, 2020. DOI: https://doi.org/10.1109/LSSC.2020.3034093.

[DT2] **D.S.. Truesdell**, S. Li, and B. Calhoun, "A 0.5V 560-kHz 18.8-fJ/cycle on-chip oscillator in 65nm CMOS with 96.1ppm/°C stability using a duty-cycled digital frequency-locked loop," *IEEE Journal of Solid-State Ciruicts (JSSC), Special Issue on VLSI Circuits Symposium (Invited Paper)*, vol. 56, no. 4, pp. 1241–1253, 2021. DOI: https://doi.org/10.1109/JSSC.2020.3048664.

[DT3] X. Chen, A. Alghaihab, Y. Shi, **D.S.. Truesdell**, B. Calhoun, and D. Wentzloff, "A crystal-less BLE transmitter with clock recovery from GFSK-modulated BLE packets," *IEEE Journal of Solid-State Ciruicts (JSSC)*, 2021. DOI: https://doi.org/10.1109/JSSC.2020.3046610.

[DT4] A. Mallick, M. Bashar, **D.S.. Truesdell**, B. Calhoun, S. Joshi, , and N. Shukla, "Using synchronized oscillators to compute the maximum independent set," *Nature Communications*, 2020. DOI: https://doi.org/10.1038/s41467-020-18445-1.

[DT5] M. K. Bashar, A. Mallick, **D.S.. Truesdell**, B. H. Calhoun, S. Joshi, and N. Shukla, "Experimental demonstration of a reconfigurable coupled oscillator platform to solve the max-cut problem," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, pp. 1–1, 2020. DOI: https://doi.org/10.1109/JXCDC.2020.3025994.

[DT6] **D.S.. Truesdell**, S. Z. Ahmed, A. W. Ghosh, and B. H. Calhoun, "Minimum-energy digital computing with steep subthreshold swing tunnel FETs," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, pp. 1–1, 2020. DOI: https://doi.org/10.1109/JXCDC.2020.3024798.

[DT7] S. Gupta, **D.S.. Truesdell**, and B. H. Calhoun, "A 65nm 16kb SRAM with 131.5pW leakage at 0.9V for wireless IoT sensor nodes," in *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2, 2020. DOI: https://doi.org/10.1109/VLSICircuits18222.2020.9162772.

[DT8] **D.S.. Truesdell**, S. Li, and B. H. Calhoun, "A 0.5V 560khz 18.8fJ/cycle ultra-low energy oscillator in 65nm CMOS with 96.1ppm/°C stability using a duty-cycled digital frequency-locked loop," in *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2, 2020. DOI: https://doi.org/10.1109/VLSICircuits18222.2020.9162832.

[DT9] P. Bassirian, D. Duvvuri, N. Liu, **D.S.. Truesdell**, H. Y. Tsao, N. S. Barker, B. H. Calhoun, and S. M. Bowers, "Design of an S-band nanowatt-level wakeup receiver with envelope detector-first architecture," *IEEE Transactions on Microwave Theory and Techniques*, vol. 68, no. 9, pp. 3920–3929, 2020. DOI: https://doi.org/10.1109/TMTT.2020.2987786.

[DT10] A. Dissanayake, J. Moody, H. L. Bishop, **D.S.**. **Truesdell**, H. Muhlbauer, R. Lu, A. Gao, S. Gong, B. H. Calhoun, and S. M. Bowers, "A- 108dbm sensitivity, -28dB SIR, 130nW to 41uW, digitally reconfigurable bit-level duty-cycled wakeup and data receiver," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–4, 2020. DOI: https://doi.org/10.1109/CICC48029.2020.9075907.

[DT11] S. Z. Ahmed, **D.S.**. **Truesdell**, Y. Tan, B. H. Calhoun, and A. W. Ghosh, "A comprehensive analysis of auger generation impacted planar tunnel FETs," *Solid-State Electronics*, vol. 169, p. 107782, 2020. DOI: https://doi.org/10.1016/j.sse.2020.107782.

[DT12] A. Alghaihab, X. Chen, Y. Shi, **D.S.**. **Truesdell**, B. H. Calhoun, and D. D. Wentzloff, "30.7 a crystal-less ble transmitter with -86dbm frequency-hopping back-channel WRX and over-the-air clock recovery from a GFSK-modulated BLE packet," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 472–474, 2020. DOI: https://doi.org/10.1109/ISSCC19947.2020.9062935.

[DT13] P. Bassirian, D. Duvvuri, **D.S.**. **Truesdell**, N. Liu, B. H. Calhoun, and S. M. Bowers, "30.1 a temperature-robust 27.6nW -65dBm wakeup receiver at 9.6GHz X-band," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 460–462, 2020. DOI: https://doi.org/10.1109/ISSCC19947.2020.9063015.

[DT14] J. Moody, A. Dissanayake, H. Bishop, R. Lu, N. Liu, D. Duvvuri, A. Gao, **D.S.**. **Truesdell**, N. S. Barker, S. Gong, B. H. Calhoun, and S. M. Bowers, "A highly reconfigurable bit-level duty-cycled trf receiver achieving -106-dBm sensitivity and 33-nW average power consumption," *IEEE Solid-State Circuits Letters*, vol. 2, no. 12, pp. 309–312, 2019. DOI: https://doi.org/10.1109/LSSC.2019.2956419.

[DT15] **D.S.**. **Truesdell**, A. Dissanayake, and B. H. Calhoun, "A 0.6-V 4.6-fJ/cycle energy-optimized frequency-locked loop in 65-nm CMOS with 20.3-ppm/℃ stability," *IEEE Solid-State Circuits Letters*, vol. 2, no. 10, pp. 223–226, 2019. DOI: https://doi.org/10.1109/LSSC.2019.2946767.

[DT16] **D.S.**. **Truesdell**, B. H. Calhoun, and S. M. Bowers, "Improving dynamic leakage suppression logic with forward body bias in 65nm CMOS," in *2019 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2019. DOI: https://doi.org/10.1109/S3S46989.2019.9320713.

[DT17] **D.S.**. **Truesdell**, J. Breiholz, S. Kamineni, N. Liu, A. Magyar, and B. H. Calhoun, "A 6–140-nW 11 Hz–8.2-kHz DVFS RISC-V microprocessor using scalable dynamic leakage-suppression logic," *IEEE Solid-State Circuits Letters*, vol. 2, no. 8, pp. 57–60, 2019. DOI: https://doi.org/10.1109/LSSC.2019.2938897.

[DT18] N. Liu, R. Agarwala, A. Dissanayake, **D.S.**. **Truesdell**, S. Kamineni, and B. H. Calhoun, "A 2.5 ppm/℃ 1.05-MHz relaxation oscillator with dynamic frequency-error compensation and fast start-up time," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1952–1959, 2019. DOI: https://doi.org/10.1109/JSSC.2019.2911208.

[DT19] J. Moody, A. Dissanayake, H. Bishop, R. Lu, N. Liu, D. Duvvuri, A. Gao, **D.S.**. **Truesdell**, N. S. Barker, S. Gong, B. H. Calhoun, and S. M. Bowers, "A -106dBm 33nW bit-level duty-cycled tuned RF wake-up receiver," in *2019 Symposium on VLSI Circuits*, pp. C86–C87, 2019. DOI: https://doi.org/10.23919/VLSIC.2019.8777956.

[DT20] **D.S.**. **Truesdell** and B. H. Calhoun, "A 640 pW 22 pJ/sample gate leakage-based digital CMOS temperature sensor with 0.25℃ resolution," in *2019 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–4, 2019. DOI: https://doi.org/10.1109/CICC.2019.8780382.

[DT21] S. Ahmed, Y. Tan, **D.S.**. **Truesdell**, B. Calhoun, and A. Ghosh, "Modeling tunnel field effect transistors–from interface chemistry to non-idealities to circuit level performance," *Journal of Applied Physics*, vol. 124, no. 15, p. 154503, 2018. DOI: https://doi.org/10.1063/1.5044434.

[DT22] **D.S.**. **Truesdell** and B. H. Calhoun, "Channel length sizing for power minimization in leakage-dominated digital circuits," in *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1–2, 2018. DOI: https://doi.org/10.1109/S3S.2018.8640174.

[DT23] N. Liu, R. Agarwala, A. Dissanayake, **D.S.**. **Truesdell**, S. Kamineni, X. Chen, D. D. Wentzloff, and B. H. Calhoun, "A 2.5 ppm/℃ 1.05 MHz relaxation oscillator with dynamic frequency-error compensation and 8ţs start-up time," in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 150–153, 2018. DOI: https://doi.org/10.1109/ESSCIRC.2018.8494338.

[DT24] S. Ahmed, Y. Tan, **D.S.**. **Truesdell**, and A. Ghosh, "Auger effect limited performance in tunnel field effect transistors," in *2017 Fifth Berkeley Symposium on Energy Efficient Electronic Systems Steep Transistors Workshop (E3S)*, pp. 1–2, 2017. DOI: https://doi.org/10.1109/E3S.2017.8246156.

B Abbreviations

ACG Adaptive Clock Generator. x, xiv, 56, 61, 64

BL Bit Line. x, 65, 66, 68, 70, 76

BLE Bluetooth Low-Energy. 5

CLK Clock. 122

CMOS Complementary Metal-Oxide Semiconductor. viii, ix, x, xi, xii, xiii, 1, 3, 5, 11, 12, 13, 15, 16, 17, 18, 19, 21, 27, 31, 37, 38, 40, 48, 51, 53, 56, 59, 60, 61, 64, 68, 72, 73, 76, 80, 98, 102, 107, 112, 129, 134, 138, 139, 140, 141

CMP Comparator. 122, 125, 126, 127, 128

DCO Digitally-controlled Oscillator. xii, xiii, 114, 115, 116, 118, 119, 120, 121, 122, 125, 126, 127, 128, 129, 131, 129, 131, 132

DFLL Digital Frequency-Locked Loop. xii, xiii, 114, 115, 116, 118, 119, 118, 120, 121, 122, 126, 127, 128, 129, 131, 129, 131, 132, 131, 132, 134, 135, 140, 141

DIBL Drain-Induced Barrier Lowering. 15, 16, 17, 20, 23, 24, 30, 35, 36, 37, 40, 53, 66, 73

DLS Dynamic Leakage Suppression. iii, v, viii, ix, x, xi, xiv, 3, 19, 20, 21, 22, 24, 25, 24, 25, 26, 27, 29, 31, 33, 34, 35, 36, 37, 38, 40, 46, 48, 51, 53, 64, 65, 68, 70, 72, 73, 76, 79, 80, 122, 137, 138, 139, 141

DRV Data Retention Voltage. x, 72, 76

DVFS Dynamic Voltage and Frequency Scaling. x, 12, 53, 56, 59, 61, 64, 66, 89

DZ Deadzone. 125, 126, 127

ESD Electrostatic Discharge. 40

FBB Forward Body Bias. ix, 24, 38, 40, 48

FLL Frequency-Locked Loop. iii, xi, xii, xiii, xiv, 4, 81, 82, 85, 86, 87, 89, 90, 103, 106, 107, 109, 110, 111, 110, 111, 112, 111, 112, 114, 115, 116, 120, 121, 122, 129, 132, 134, 135, 140, 141

FVC Frequency-to-Voltage Converter. xi, xii, 86, 87, 88, 89, 103, 106, 115, 116, 117, 118, 120, 122, 125, 126, 127, 128

GIDL Gate-Induced Drain Leakage. viii, 14, 15, 16, 24

GPIO General-Purpose Input/Output. x, 1, 59, 61

HVT High Threshold Voltage. ix, 24, 38, 40, 48, 56, 67, 76, 98, 138

I/O Input/Output. iii, 2, 38, 40, 68, 70, 76

IoT Internet of Things. iii, 1, 2, 5, 7, 8, 12, 65, 66, 70, 81, 82, 111, 120, 137

kb Kilobit. xi, 76, 80, 139

kB Kilobyte. 5, 65, 66

KCL Kirchoff's Current Law. 22, 31

LED Light-Emitting Diode. 5

LF Low-Pass Filter. 88, 103

LSB Least-Significant Bit. 125

LVT Low Threshold Voltage. 24, 38, 53, 56, 76, 98

MAC Multiply-Accumulate. ix, 3, 40, 48

MOSFET Metal-Oxide-Semiconductor Field-Effect Transistor. 13, 14

MSB Most-Significant Bit. 98

NM Noise Margin. 27, 38, 40, 46, 65, 66, 68, 72

NVM Non-Volatile Memory. 7

PD Pull-Down. xi, 26, 36, 40, 73

PLL Phase-Locked Loop. xi, 81, 85, 87

PMU Power Management Unit. 7

ppm Parts-per-Million. 98, 102, 104, 109, 110, 111, 112, 118, 119, 122, 129, 132, 134

PU Pull-Up. xi, 23, 25, 26, 40, 70, 73

RBB Reverse Body Bias. 15, 38, 67

RBL Read Bit Line. 70

RF Radio Frequency. 1

RO Ring Oscillator. x, 56, 81, 103, 106, 122

RWL Read Word Line. 70

RX receive. 8

RXO Relaxation Oscillator. xi, xii, xiii, xiv, 4, 81, 82, 83, 84, 85, 87, 89, 90, 91, 96, 98, 101, 102, 134, 135, 139, 140

SAR Successive-Approximation Register. 114, 122, 125

SDLS Scalable Dynamic Leakage Suppression. ix, x, xiv, 53, 56, 59, 60, 61, 64, 61, 64, 138

SIP System-in-Package. 5

SM Switching Margin. 27, 38, 40, 46, 72

SoC System-on-Chip. 7, 70, 120

SPI Serial-Peripheral Interface. 1, 59

SRAM Static Random-Access Memory. iii, x, xi, xiv, 1, 3, 5, 19, 59, 65, 66, 67, 68, 70, 72, 73, 76, 79, 80, 138, 139

SS Steady-State. 126

TC Temperature Coefficient. xii, xiii, 98, 102, 105, 107, 110, 111, 110, 111, 116, 119, 122, 129, 134, 135

TX transmit. 8

VCO Voltage-Controlled Oscillator. xii, 87, 88, 89, 90, 103, 106, 107, 114, 115, 129

VD Voltage Detector. 86, 89, 90, 103, 105, 106, 110, 114, 115

VSC Voltage Scaling Controller. x, xiv, 56, 59, 61, 64

VTC Voltage Transfer Curve. ix, 26, 31, 33, 40, 46, 65, 73, xxiv

WBL Write Bit Line. 70

WL Word Line. x, 65, 66, 68, 70, 73, 76, 139

WWL Write Word Line. 70

XO Crystal Oscillator. 81, 141

C Algorithms

Algorithm 1 (MATLAB) Input threshold voltage V_{ih} and V_{il} for an inverter VTC

```
1:  for x1_iterable=1:length(VIN_values_0_1) do
2:      x2=0;
3:      x1=VIN_values_0_1(x1_iterable);
4:      y1=VOUT_values_0_1(x1_iterable);
5:      for  x1_sub_iterable=1:length(VIN_values_0_1)  do
6:          if  VOUT_values_1_0(x1_sub_iterable)>=x1 then
7:              x2=VIN_values_1_0(x1_sub_iterable);
8:              y2=VOUT_values_1_0(x1_sub_iterable);
9:              break
10:     if x2>=y1 and x1>=x2 then
11:         VINH=x1;
12:         VINL=x2;
13:         break
```

Algorithm 2 (MATLAB) High logic output voltage V_{oh} for an inverter VTC

```
1:  for  x1_iterable=1:length(VIN_values_0_1) do
2:      x1=VIN_values_0_1(x1_iterable);
3:      y1=VOUT_values_0_1(x1_iterable);
4:      [a,b]=min(abs(VIN_values_0_1-y1));
5:      x2_index=b;
6:      x2=VIN_values_0_1(x2_index);
7:      k=VOUT_values_0_1(x2_index)-x1;
8:      if  k<=0 then
9:          VOH=y1;
10:         break
```

Algorithm 3 (MATLAB) Low logic output voltage V_{ol} for an inverter VTC

```
1:  for  x1_iterable=length(VIN_values_0_1):-1:1 do
2:      x1=VIN_values_0_1(x1_iterable);
3:      y1=VOUT_values_0_1(x1_iterable);
4:      [a,b]=min(abs(VIN_values_0_1-y1));
5:      x2_index=b;
6:      x2=VIN_values_0_1(x2_index);
7:      k=VOUT_values_0_1(x2_index)-x1;
8:      if  k>=0 then
9:          VOL=y1;
10:         break
```

D References

[1] W. Lim, I. Lee, D. Sylvester, and D. Blaauw, "8.2 batteryless sub-nw cortex-m0+ processor with dynamic leakage-suppression logic," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pp. 1–3, 2015.

[2] L. Lin, S. Jain, and M. Alioto, "A 595pw 14pj/cycle microcontroller with dual-mode standard cells and self-startup for battery-indifferent distributed sensing," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pp. 44–46, 2018.

[3] S. Hanson, M. Seok, Y. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "A low-voltage processor for sensing applications with picowatt standby mode," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1145–1155, 2009.

[4] 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.

[5] D. Kim, G. Chen, M. Fojtik, M. Seok, D. Blaauw, and D. Sylvester, "A 1.85fw/bit ultra low leakage 10t sram with speed compensation scheme," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pp. 69–72, 2011.

[6] J. Breiholz, *Digital Circuits and Systems for Ultra-Low Power Internet of Things Applications*. PhD thesis, 2021.

[7] P. Newswire and Verified Market Research, "Internet of things (iot) market worth $1319.08 billion, globally, by 2026 at 25.68% cagr: Verified market research." Available at https://www.prnewswire.com.

[8] Everactive, "The battery problem: An infographic." Available at https://everactive.com/resources/.

[9] I. Insights, "Mcus sales to reach record-high annual revenues through 2022." Available at https://www.icinsights.com/news/bulletins/.

[10] I. Insights, "Microcontrollers will regain growth after 2019 slump." Available at https://www.icinsights.com/news/bulletins/.

[11] S. Kim, R. Vyas, J. Bito, K. Niotaki, A. Collado, A. Georgiadis, and M. M. Tentzeris, "Ambient rf energy-harvesting technologies for self-sustainable standalone wireless sensor platforms," *Proceedings of the IEEE*, vol. 102, no. 11, pp. 1649–1666, 2014.

[12] M. Piñuela, P. D. Mitcheson, and S. Lucyszyn, "Ambient rf energy harvesting in urban and semi-urban environments," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 7, pp. 2715–2726, 2013.

[13] X. Liu, H. Gao, J. E. Ward, X. Liu, B. Yin, T. Fu, J. Chen, D. R. Lovley, and J. Yao, "Power generation from ambient humidity using protein nanowires," *Nature*, vol. 578, pp. 550–554, Feb 2020.

[14] S. Bandyopadhyay, P. P. Mercier, A. C. Lysaght, K. M. Stankovic, and A. P. Chandrakasan, "A 1.1 nw energy-harvesting system with 544 pw quiescent power for next-generation implants," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 2812–2824, 2014.

[15] A. J. Bandodkar, J.-M. You, N.-H. Kim, Y. Gu, R. Kumar, A. M. V. Mohan, J. Kurniawan, S. Imani, T. Nakagawa, B. Parish, M. Parthasarathy, P. P. Mercier, S. Xu, and J. Wang, "Soft, stretchable, high power density electronic skin-based biofuel cells for scavenging energy from human sweat," *Energy Environ. Sci.*, vol. 10, pp. 1581–1589, 2017.

[16] B. J. Hansen, Y. Liu, R. Yang, and Z. L. Wang, "Hybrid nanogenerator for concurrently harvesting biomechanical and biochemical energy," *ACS Nano*, vol. 4, no. 7, pp. 3647–3652, 2010. PMID: 20507155.

[17] P. M. Thibado, P. Kumar, S. Singh, M. Ruiz-Garcia, A. Lasanta, and L. L. Bonilla, "Fluctuation-induced current from freestanding graphene," *Phys. Rev. E*, vol. 102, p. 042101, Oct 2020.

[18] Espressif Systems, "Espressif systems esp32-pico-d4." Available at https://www.espressif.com/.

[19] T. Instruments, "Ulp meets energy harvesting: a game-changing combination for design engineers." Available at https://www.mouser.ph/pdfDocs/TI-ULP-meets-energy-harvesting-A-game-changing-combination-for-design-engineers.pdf.

[20] E. E. Aktakka and K. Najafi, "A micro inertial energy harvesting platform with self-supplied power management circuit for autonomous wireless sensor nodes," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 9, pp. 2017–2029, 2014.

[21] J. K. Brown, D. Abdallah, J. Boley, N. Collins, K. Craig, G. Glennon, K. Huang, C. J. Lukas, W. Moore, R. K. Sawyer, Y. Shakhsheer, F. B. Yahya, A. Wang, N. E. Roberts, D. D. Wentzloff, and B. H. Calhoun, "27.1 a 65nm energy-harvesting ulp soc with 256kb cortex-m0 enabling an 89.1μw continuous machine health monitoring wireless self-powered system," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 420–422, 2020.

[22] S. Bose, B. Shen, and M. L. Johnston, "A batteryless motion-adaptive heartbeat detection system-on-chip powered by human body heat," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 11, pp. 2902–2913, 2020.

[23] L. Lin, S. Jain, and M. Alioto, "Sub-nw microcontroller with dual-mode logic and self-startup for battery-indifferent sensor nodes," *IEEE Journal of Solid-State Circuits*, pp. 1–1, 2020.

[24] F. Laman and K. Brandt, "Effect of discharge current on cycle life of a rechargeable lithium battery," *Journal of Power Sources*, vol. 24, no. 3, pp. 195–206, 1988.

[25] C. J. Lukas, F. B. Yahya, J. Breiholz, A. Roy, X. Chen, H. N. Patel, N. Liu, A. Kosari, S. Li, D. Akella Kamakshi, O. Ayorinde, D. D. Wentzloff, and B. H. Calhoun, "A 1.02 μw battery-less, continuous sensing and post-processing sip for wearable applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 2, pp. 271–281, 2019.

[26] B. H. Calhoun, S. Khanna, Y. Zhang, J. Ryan, and B. Otis, “System design principles combining sub-threshold circuit and architectures with energy scavenging mechanisms,” in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 269–272, 2010.

[27] P. P. Mercier, S. Bandyopadhyay, A. C. Lysaght, K. M. Stankovic, and A. P. Chandrakasan, “A sub-nw 2.4 ghz transmitter for low data-rate sensing applications,” *IEEE Journal of Solid-State Circuits*, vol. 49, no. 7, pp. 1463–1474, 2014.

[28] R. Zimmermann and W. Fichtner, “Low-power logic styles: Cmos versus pass-transistor logic,” *IEEE Journal of Solid-State Circuits*, vol. 32, no. 7, pp. 1079–1090, 1997.

[29] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, “Low-power cmos digital design,” *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, 1992.

[30] E. Keonjian, “Micropower electronics,” *Pergamon Press*, 1964.

[31] F. Wanlass and C. Sah, “Nanowatt logic using field-effect metal-oxide semiconductor triodes,” in *1963 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, vol. VI, pp. 32–33, 1963.

[32] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, “A dynamic voltage scaled microprocessor system,” *IEEE Journal of Solid-State Circuits*, vol. 35, no. 11, pp. 1571–1580, 2000.

[33] V. Gutnik and A. P. Chandrakasan, “Embedded power supply for low-power dsp,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 5, no. 4, pp. 425–435, 1997.

[34] A. Chaudhry, “Fundamentals of nanoscaled field effect transistors„” *Springer*, 2013.

[35] J. C. Ranuárez, M. Deen, and C.-H. Chen, “A review of gate tunneling current in mos devices,” *Microelectronics Reliability*, vol. 46, no. 12, pp. 1939–1956, 2006.

[36] K. M. Cao, W. . Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu, “Bsim4 gate leakage model including source-drain partition,” in *International Electron Devices Meeting 2000. Technical Digest. IEDM (Cat. No.00CH37138)*, pp. 815–818, 2000.

[37] Berkeley BSIM Group, “Bsim4.8 model,” *Online (open-source)*, 2013. Available at https://bsim.berkeley.edu/models/bsim4/.

[38] B. H. Calhoun, “Low energy digital circuit design using sub-threshold operation,” *Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science*, 2006.

[39] L. Benini, P. Siegel, and G. De Micheli, “Saving power by synthesizing gated clocks for sequential circuits,” *IEEE Design Test of Computers*, vol. 11, no. 4, pp. 32–41, 1994.

[40] Qing Wu, M. Pedram, and Xunwei Wu, “Clock-gating and its application to low power design of sequential circuits,” *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 3, pp. 415–420, 2000.

[41] J. D. Meindl and J. A. Davis, “The fundamental limit on binary switching energy for terascale integration (tsi),” *IEEE Journal of Solid-State Circuits*, vol. 35, no. 10, pp. 1515–1516, 2000.

[42] R. Landauer, “Irreversibility and heat generation in the computing process,” *IBM Journal of Research and Development*, vol. 5, no. 3, pp. 183–191, 1961.

[43] R. W. Keyes, “Physical limits in digital electronics,” *Proceedings of the IEEE*, vol. 63, no. 5, pp. 740–767, 1975.

[44] N. Lotze and Y. Manoli, “A 62 mv 0.13 μ m cmos standard-cell-based design technique using schmitt-trigger logic,” *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 47–60, 2012.

[45] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, “Scaling of stack effect and its application for leakage reduction,” in *ISLPED’01: Proceedings of the 2001 International Symposium on Low Power Electronics and Design (IEEE Cat. No.01TH8581)*, pp. 195–200, 2001.

[46] Y. Ye, S. Borkar, and V. De, “A new technique for standby leakage reduction in high-performance circuits,” in *1998 Symposium on VLSI Circuits. Digest of Technical Papers (Cat. No.98CH36215)*, pp. 40–41, 1998.

[47] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, “1-v power supply high-speed digital circuit technology with multithreshold-voltage cmos,” *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, 1995.

[48] K. Shi and D. Howard, “Challenges in sleep transistor design and implementation in low-power designs,” in *Proceedings of the 43rd Annual Design Automation Conference*, DAC ’06, (New York, NY, USA), pp. 113–116, Association for Computing Machinery, 2006.

[49] S. Sankar, M. Goel, P. H. Chen, V. R. Rao, and M. S. Baghini, “Switched-capacitor-assisted power gating for ultra-low standby power in cmos digital ics,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4281–4294, 2020.

[50] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, “A 0.9-v, 150-mhz, 10-mw, 4 mm/sup 2/, 2-d discrete cosine transform core processor with variable threshold-voltage (vt) scheme,” *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1770–1779, 1996.

[51] N. Sirisantana, L. Wei, and K. Roy, “High-performance low-power cmos circuits using multiple channel length and multiple oxide thickness,” in *Proceedings 2000 International Conference on Computer Design*, pp. 227–232, 2000.

[52] P. Gupta, A. B. Kahng, P. Sharma, and D. Sylvester, “Gate-length biasing for runtime-leakage control,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 8, pp. 1475–1485, 2006.

[53] D. Levacq, V. Dessard, and D. Flandre, “Low leakage soi cmos static memory cell with ultra-low power diode,” *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 689–702, 2007.

[54] D. Bol, R. Ambroise, D. Flandre, and J. Legat, “Building ultra-low-power low-frequency digital circuits with high-speed devices,” in *2007 14th IEEE International Conference on Electronics, Circuits and Systems*, pp. 1404–1407, 2007.

[55] D. Bol, “Pushing ultra-low-power digital circuits into the nanometer era,” *UC Louvain (dissertation)*, 2008.

[56] D. Bol, J. De Vos, R. Ambroise, D. Flandre, and J.-D. Legat, “Building ultra-low-power high-temperature digital circuits in standard high-performance soi technology,” *Solid-State Electronics*, vol. 52, no. 12, pp. 1939–1945, 2008. Selected Papers from the EUROSOI ’08 Conference.

[57] O. Aiello, P. Crovetti, L. Lin, and M. Alioto, “A pw-power hz-range oscillator operating with a 0.3-1.8-v unregulated supply,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 5, pp. 1487–1496, 2019.

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

[59] K. A. et al., “The rocket chip generator,” *Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley*, vol. Tech. Rep. UCB/EECS-2016-17, 2011.

[60] R. Salvador, A. Sanchez, X. Fan, and T. Gemmeke, “A cortex-m3 based mcu featuring avs with 34nw static power, 15.3pj/inst. active energy, and 16power variation across process and temperature,” in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 278–281, 2018.

[61] E. Seevinck, F. J. List, and J. Lohstroh, “Static-noise margin analysis of mos sram cells,” *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, 1987.

[62] A. J. Bhavnagarwala, S. V. Kosonocky, S. P. Kowalczyk, R. V. Joshi, Y. H. Chan, U. Srinivasan, and J. K. Wadhwa, “A transregional cmos sram with single, logic VDD and dynamic power rails,” in *2004 Symposium on VLSI Circuits. Digest of Technical Papers (IEEE Cat. No.04CH37525)*, pp. 292–293, 2004.

[63] K. Kanda, T. Miyazaki, Min Kyeong Sik, H. Kawaguchi, and T. Sakurai, “Two orders of magnitude leakage power reduction of low voltage srams by row-by-row dynamic VDD control (rrdv) scheme,” in *15th Annual IEEE International ASIC/SOC Conference*, pp. 381–385, 2002.

[64] B. H. Calhoun and A. P. Chandrakasan, “Static noise margin variation for sub-threshold sram in 65-nm cmos,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, 2006.

[65] B. H. Calhoun and A. P. Chandrakasan, “A 256-kb 65-nm sub-threshold sram design for ultra-low-voltage operation,” *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, 2007.

[66] J. P. Kulkarni, K. Kim, and K. Roy, “A 160 mv robust schmitt trigger based subthreshold sram,” *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, 2007.

[67] 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.

[68] N. Verma and A. P. Chandrakasan, “A 256 kb 65 nm 8t subthreshold sram employing sense-amplifier redundancy,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, 2008.

[69] J. Lee, M. Saligane, D. Blaauw, and D. Sylvester, “A 0.3-v to 1.8âĂŞ3.3-v leakage-biased synchronous level converter for ulp socs,” *IEEE Solid-State Circuits Letters*, vol. 3, pp. 130–133, 2020.

[70] H. You, J. Yuan, W. Tang, S. Qiao, and Y. Hei, “An energy-efficient level shifter for ultra low-voltage digital lsis,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3357–3361, 2020.

[71] M. Lanuzza, F. Crupi, S. Rao, R. De Rose, S. Strangio, and G. Iannaccone, “An ultralow-voltage energy-efficient level shifter,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 1, pp. 61–65, 2017.

[72] J. Lim, T. Jang, M. Saligane, M. Yasuda, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, “A 224 pw 260 ppm/℃ gate-leakage-based timer for ultra-low power sensor nodes with second-order temperature dependency cancellation,” in *2018 IEEE Symposium on VLSI Circuits*, pp. 117–118, 2018.

[73] X. Zhang and A. B. Apsel, “A low-power, process-and- temperature- compensated ring oscillator with addition-based current source,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 5, pp. 868–878, 2011.

[74] M. P. Flynn and S. U. Lidholm, “A 1.2- mu m cmos current-controlled oscillator,” *IEEE Journal of Solid-State Circuits*, vol. 27, no. 7, pp. 982–987, 1992.

[75] S. Dai and J. K. Rosenstein, “A 14.4nw 122khz dual-phase current-mode relaxation oscillator for near-zero-power sensors,” in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–4, 2015.

[76] A. Savanth, A. S. Weddell, J. Myers, D. Flynn, and B. M. Al-Hashimi, “A sub-nw/khz relaxation oscillator with ratioed reference and sub-clock power gated comparator,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3097–3106, 2019.

[77] H. Jiang, P. P. Wang, P. P. Mercier, and D. A. Hall, “A 0.4-v 0.93-nw/khz relaxation oscillator exploiting comparator temperature-dependent delay to achieve 94-ppm/Âřc stability,” *IEEE Journal of Solid-State Circuits*, vol. 53, no. 10, pp. 3004–3011, 2018.

[78] M. Choi, T. Jang, S. Bang, Y. Shi, D. Blaauw, and D. Sylvester, “A 110 nw resistive frequency locked on-chip oscillator with 34.3 ppm/℃ temperature stability for system-on-chip designs,” *IEEE Journal of Solid-State Circuits*, vol. 51, no. 9, pp. 2106–2118, 2016.

[79] J. Jung, I. Kim, S. Kim, Y. Lee, and J. Chun, “A 1.08-nw/khz 13.2-ppm/℃ self-biased timer using temperature-insensitive resistive current,” *IEEE Journal of Solid-State Circuits*, vol. 53, no. 8, pp. 2311–2318, 2018.

[80] M. Ding, Z. Zhou, S. Traferro, Y. Liu, C. Bachmann, and F. Sebastiano, “A 33-ppm/℃ 240-nw 40-nm cmos wakeup timer based on a bang-bang digital-intensive frequency-locked-loop for iot applications,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 7, pp. 2263–2273, 2020.

[81] K. Ueno, T. Asai, and Y. Amemiya, “A 30-mhz, 90-ppm/℃ fully-integrated clock reference generator with frequency-locked loop,” in *2009 Proceedings of ESSCIRC*, pp. 392–395, 2009.

[82] M. Ding, M. Song, E. Tiurin, S. Traferro, Y. H. Liu, and C. Bachmann, “A 0.9pj/cycle 8ppm/℃ dfll-based wakeup timer enabled by a time-domain trimming and an embedded temperature sensing,” in *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2, 2020.

[83] C. Gurleyuk, L. Pedala, S. Pan, F. Sebastiano, and K. A. A. Makinwa, “A cmos dual-rc frequency reference with Âś200-ppm inaccuracy from âĹŠ45℃ to 85℃,” *IEEE Journal of Solid-State Circuits*, vol. 53, no. 12, pp. 3386–3395, 2018.

[84] G. Cristiano, J. Liao, A. Novello, G. Atzeni, and T. Jang, “A 8.7ppm/℃, 694nw, one-point calibrated rc oscillator using a nonlinearity-aware dual phase-locked loop and dsm-controlled frequency-locked loops,” in *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2, 2020.

[85] U. Denier, “Analysis and design of an ultralow-power cmos relaxation oscillator,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 8, pp. 1973–1982, 2010.

[86] A. Paidimarri, D. Griffith, A. Wang, G. Burra, and A. P. Chandrakasan, “An rc oscillator with comparator offset cancellation,” *IEEE Journal of Solid-State Circuits*, vol. 51, no. 8, pp. 1866–1877, 2016.

[87] C. W. Malinowski, H. Rinderle, and M. Siegle, “A novel frequency-processing method and its implications on future tuning systems,” *IEEE Transactions on Consumer Electronics*, vol. CE-25, no. 4, pp. 649–669, 1979.

[88] K. Clarke and D. Hess, “Frequency locked loop fm demodulator,” *IEEE Transactions on Communication Technology*, vol. 15, no. 4, pp. 518–524, 1967.

[89] H. T. Bui and Y. Savaria, “Design of a high-speed differential frequency-to-voltage converter and its application in a 5-ghz frequency-locked loop,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 3, pp. 766–774, 2008.

[90] F. Gardner, “Properties of frequency difference detectors,” *IEEE Transactions on Communications*, vol. 33, no. 2, pp. 131–138, 1985.

[91] J. A. Andreas Michaelsen and D. T. Wisland, “A low-voltage low-power frequency-to-voltage converter for vco feedback linearization,” in *2010 17th IEEE International Conference on Electronics, Circuits and Systems*, pp. 1132–1135, 2010.

[92] C. Gurleyuk, B. Pak, and D. Y. Aksin, “A frequency to voltage converter based on an accurate pulse width modulator for frequency locked loops,” in *2013 IEEE 4th Latin American Symposium on Circuits and Systems (LASCAS)*, pp. 1–4, 2013.

[93] T. . Liu and R. G. Meyer, "A 250-mhz monolithic voltage-controlled oscillator with low temperature coefficient," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 555–561, 1990.

[94] A. Djemouai, M. A. Sawan, and M. Slamani, "New frequency-locked loop based on cmos frequency-to-voltage converter: design and implementation," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 48, no. 5, pp. 441–449, 2001.

[95] F. u. Rahman, S. Kim, N. John, R. Kumar, X. Li, R. Pamula, K. A. Bowman, and V. S. Sathe, "A unified clock and switched-capacitor-based power delivery architecture for variation tolerance in low-voltage soc domains," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1173–1184, 2019.

[96] W. Riley, "Handbook of frequency stability analysis," *NIST Special Publication 1065*, 2008. Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50505.

[97] J. Lee, G. Bosman, K. R. Green, and D. Ladwig, "Noise model of gate-leakage current in ultrathin oxide mosfets," *IEEE Transactions on Electron Devices*, vol. 50, no. 12, pp. 2499–2506, 2003.

[98] M. Ding, Z. Zhou, Y. Liu, S. Traferro, C. Bachmann, K. Philips, and F. Sebastiano, "A 0.7-v 0.43-pj/cycle wakeup timer based on a bang-bang digital-intensive frequency-locked-loop for iot applications," *IEEE Solid-State Circuits Letters*, vol. 1, no. 2, pp. 30–33, 2018.

[99] T. Jang, M. Choi, S. Jeong, S. Bang, D. Sylvester, and D. Blaauw, "5.8 a 4.7nw 13.8ppm/℃ self-biased wakeup timer using a switched-resistor scheme," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 102–103, 2016.

[100] T. Jang, S. Jeong, D. Jeon, K. D. Choo, D. Sylvester, and D. Blaauw, "A noise reconfigurable all-digital phase-locked loop using a switched capacitor-based frequency-locked loop and a noise detector," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 50–65, 2018.

[101] M. Seok, G. Kim, D. Blaauw, and D. Sylvester, "A portable 2-transistor picowatt temperature-compensated voltage reference operating at 0.5 v," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 10, pp. 2534–2545, 2012.

[102] X. Wu, I. Lee, Q. Dong, K. Yang, D. Kim, J. Wang, Y. Peng, Y. Zhang, M. Saligane, M. Yasuda, K. Kumeno, F. Ohno, S. Miyoshi, M. Kawaminami, D. Sylvester, and D. Blaauw, "A 0.04mm^3 16nw wireless and batteryless sensor system with integrated cortex-m0+ processor and optical communication for cellular temperature measurement," in *2018 IEEE Symposium on VLSI Circuits*, pp. 191–192, 2018.

[103] L. Lin, S. Jain, and M. Alioto, "Multi-sensor platform with five-order-of-magnitude system power adaptation down to 3.1nw and sustained operation under moonlight harvesting," in *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2, 2020.

[104] S. Jeong, I. Lee, D. Blaauw, and D. Sylvester, "A 5.8 nw cmos wake-up timer for ultra-low-power wireless applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 8, pp. 1754–1763, 2015.