2018

Integrated Circuits for Programming Flash Memories in Portable Applications

Mir Mohammad Navidi

Integrated Circuits for Programming Flash Memories in Portable Applications

Mir Mohammad Navidi

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

David W. Graham, Ph.D., Chair
Vinod Kulathumani, Ph.D.
Parviz Famouri, Ph.D.
Jeremy Dawson, Ph.D.
Edward M. Sabolsky, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2018

Keywords: Analog, Integrated Circuits, Floating-Gate, Flash Memory, Field Programmable Analog Arrays

Abstract

Integrated Circuits for Programming Flash Memories in Portable Applications

Mir Mohammad Navidi

Smart devices such as smart grids, smart home devices, etc. are infrastructure systems that connect the world around us more than before. These devices can communicate with each other and help us manage our environment. This concept is called the Internet of Things (IoT). Not many smart nodes exist that are both low-power and programmable. Floating-gate (FG) transistors could be used to create adaptive sensor nodes by providing programmable bias currents. FG transistors are mostly used in digital applications like Flash memories. However, FG transistors can be used in analog applications, too. Unfortunately, due to the expensive infrastructure required for programming these transistors, they have not been economical to be used in portable applications. In this work, we present low-power approaches to programming FG transistors which make them a good candidate to be employed in future wireless sensor nodes and portable systems. First, we focus on the design of low-power circuits which can be used in programming the FG transistors such as high-voltage charge pumps, low-drop-out regulators, and voltage reference cells. Then, to achieve the goal of reducing the power consumption in programmable sensor nodes and reducing the programming infrastructure, we present a method to program FG transistors using negative voltages. We also present charge-pump structures to generate the necessary negative voltages for programming in this new configuration.

Dedication

To my advisor and all CESLAB members who have been very helpful.

To my family and the memory of my father.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Smart devices such as smart grids, smart homes, etc. are infrastructure systems that connect the world around us more than before. A common concept of such systems is that all the smart devices are able to communicate and transmit their information using smart sensors. This concept is called the Internet of Things (IoT). In these systems, monitoring of the system is possible using a network of embedded devices. However, not many smart nodes exist that are both low-power and programmable. Analog memories like Floating-gate (FG) transistors are good candidates to create low-power and programmable nodes.

FG transistors are mostly used in digital applications like Flash memories. However, FG transistors can be used in analog applications, too. For example, FG transistors have already been used to provide programmable currents for adaptive filters in analog signal processing applications. Unfortunately, due to the expensive infrastructure required for programming FG transistors, they have not been economical to be used in portable applications. In this work, we present low-power circuits for programming FG transistors which make them a good candidate to be employed in future wireless sensor nodes and portable systems.

In order to make large-scale programmable systems using FG transistors, low-power programming circuits are required. We will show that to change the stored analog data, Fowler-Nordheim tunneling and hot-electron injection are used. Both of these techniques require high DC voltages with very small ripple. Step-up voltage converters can be used for these two procedures. Thus, we present the design procedure and experimental results of two charge pump circuits for FG programming applications. First, we focus on the design of a high-voltage charge pump for the tunneling process. In order to reduce the power consumption of the proposed charge pump for battery-operated applications, a variable-frequency regulation technique is used. A new method

for reducing start-up power consumption in a previously reported charge-transfer stage is presented. Additionally, a new circuit is presented to minimize the short-circuit current of the clock circuitry.

A second charge-pump is used for injection of FG transistors. The injection accuracy of the FG transistors is directly related to the characteristics of the high voltage signal used in the injection process. Thus a low-ripple and high-voltage signal must be employed for this process. A new low-dropout regulator (LDO) is used inside the tunneling charge-pump loop to reduce the output ripple for the injection process. A cascode structure is used to improve the PSRR of the LDO regulator. This technique relaxes the stability requirements of the LDO because a smaller pass transistor can be used in this new structure. Thus, a smaller parasitic capacitance is created at the gate of the pass transistor. One challenge of this design is to provide a bias voltage at the gate of the cascode transistor. We use the feedback resistive divider to provide a bias voltage to bias the cascode transistor in the above-threshold saturation region. To reduce the voltage overshoot in the output of the injection charge-pump, a feed-forward compensation technique is used. The resulting charge pump generates high voltages for the injection process under different load currents with a critically damped behavior. These two charge pumps can be used in other applications like MEMS applications, electret earphones, etc.

Any System-on-Chip (SoC) will require supporting circuits to provide stable DC voltages. Voltage reference cells are used to provide a stable voltage across a wide range of temperature and supply variations. We present the design procedure and experimental results of two voltage reference cells which can be used in SoC applications. First, we present a voltage reference for low-power applications with an above-$1V$ output. Thick oxide transistors are used to generate a complementary to absolute temperature (CTAT) voltage in this circuit. We employ a new curvature compensation technique in the second voltage reference cell to achieve a low temperature coefficient (TC) output voltage. Finally, we present the experimental results of two voltage regulators which generate V_{DD} and V_{mid} for analog and digital circuits. The fast transient voltage regulator is used to generate a stable supply voltage (V_{DD}) across a wide range of load currents. The push-pull voltage regulator is used to generate a stable mid-rail voltage (V_{mid}) which can provide a few microamps of sink/source load currents.

Finally, a new programming technique for analog applications is presented. We call this technique below-ground programming of FG transistors. Using negative voltages to program digital flash memories is widely used. To the best of our knowledge, storing analog data onto an array

of FG transistors in a standard CMOS process using negative voltages was not possible before. This technique will improve the FG programming accuracy and circumvents the requirement of the multiplexing circuits (e.g. high side switches). The main obstacle to this technique is to select a target FG transistor among an array of FG transistors. Since this technique requires digital circuits operating below ground. Operating digital logic below ground is not possible in a standard CMOS process because the parasitic PN junctions of the NMOS transistors are forward biased when a negative voltage is applied to the input terminals of the logic circuits. A new technique is presented to select a target FG transistor in an array of FG transistors for large-scale applications (e.g. FPAAs). Another challenge of the below-ground programming technique is designing a charge-pump that can generate a negative voltage in a standard CMOS process. The main challenge is to provide the appropriate well voltage for charge-transfer switches. These negative charge pumps can be used in other applications like TFT display drivers, etc.

This dissertation is organized into six chapters. In Chapter 2, I discuss the background of FG transistors first; then I review the conventional methods of programming the FG transistors. In Chapter 3, I explain the charge pumps required for tunneling and injection of the FG transistors. Measurement results of these charge pumps are presented too. In Chapter 4, the background of low-power voltage reference cells will be discussed first. Then, I present a low-power voltage reference cell suitable for portable applications. In Chapter 5, basic characteristics of the voltage regulators are covered first. Then, design, simulations, and measurements of two voltage regulators are presented. In Chapter 6, a below-ground charge pump is presented first. Then, a programmer circuit that uses this negative charge pump is discussed. Then, below-ground programming of the FG transistors is presented. In Appendix A, a low power voltage reference cell using a new curvature compensation technique is presented. Finally, in Appendix B a very simple subthreshold voltage reference cell is presented.

Chapter 2

Overview of Floating-Gate Transistors

Floating gate (FG) transistors have a wide range of applications from digital flash memories to programmable analog bias current generators. Different methods are used to modify the charge stored on the FG transistor. In this chapter, we will focus on the background of the FG transistors and their wide range of applications.

2.1 Flash Memory

There are two major types of CMOS memories available. First, volatile memories like random access memories (RAMs like SRAM or DRAM) lose stored information when the power supply is turned off, although they are very fast in writing and reading (SRAM) or very dense (DRAM). Second, non-volatile memories (NVM) can keep the data stored when the power supply is turned off (ROMs like EPROM, EEPROM, and Flash). There has always been a continuous growth of nonvolatile memories in the past few years, especially for Flash-memory applications.

Flash-memories have two major applications [7]. The first application is mainly integrating NVM in logic systems (e.g. microprocessors) to allow software updates, store identification codes, or reconfigure the system in the field. The second application is to create storage devices, like solid-state hard drives, composed of Flash memory arrays which are arranged to make large-size memories to replace miniaturized hard disks. Solid-state disks made by Flash-memories are suitable for portable applications because they have small dimensions and very low power consumption. A good point about Flash memories is that they combine the capability of NVM storage with an access time comparable to DRAM's. These two properties allow direct execution of microcodes in Flash-memories.

In order to have a memory cell that can switch between one state to other states and that can store the information independent of external conditions, the storing element must be a device whose conductivity changes in a non-destructive way [7]. One possible solution is to have a transistor with an adaptive threshold voltage based on the two states of the memory cell (e.g. binary value "1" and "0"). Reading a single bit is performed by applying a gate voltage that is between the values of the thresholds of the erased and programmed cells and measuring the current going through the transistor. The threshold voltage of a transistor can be expressed as:

$$V_T = K - Q/C_{ox} \tag{2.1}$$

where K is a technology-dependent parameter, Q is the charge weighted with respect to its position in the gate oxide, and C_{ox} is the gate oxide capacitance. Therefore, the threshold voltage of the memory cell can be changed by modifying the amount of charge present between the gate and channel (e.g. Q/C_{ox}). The two common ways of shifting the threshold voltage are:

1. Storing the charge in traps that are present in the oxide, or more precisely at the interface between two dielectric materials. The most commonly used interface is the silicon oxide/nitride interface. Any device that is fabricated in this way is called Metal-Nitride-Oxide-Silicon (MNOS) cells [8, 9].

2. Storing the charge in a conductive material layer between the gate and the channel which is completely surrounded by an insulator. This device is called a floating-gate (FG) transistor.

MNOS devices are not attractive in consumer electronics due to their low endurance (capability of maintaining the stored information after erase/program/read cycling) and retention (capability of keeping the stored information in time). FG devices are still the basis of every modern NVM, especially for Flash memories. In the following sections, a detailed discussion of FG-based Flash memories is presented.

2.2 Floating-Gate Transistors

The cross-section of a typical CMOS transistor and an FG transistor are shown in Fig. 2.1 and 2.2. A typical transistor (MOSFET) uses a Metal connection to change the potential at the gate of the CMOS transistor to open or close the connection between two other electrodes, called

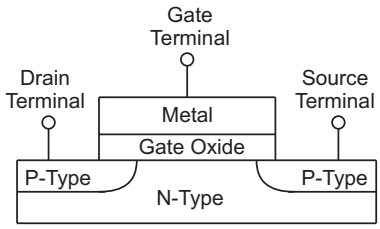


Figure 2.1: Cross section view of a typical CMOS transistor

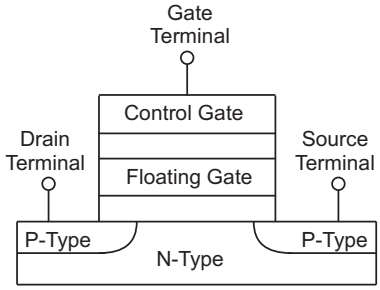


Figure 2.2: Cross section view of an FG-transistor

the source and drain, allowing the device to function as an on/off switch. When enough charge is placed onto the gate electrode, the semiconducting region becomes more conductive, and a channel will be created between the source and drain. When the charge is removed from the gate electrode, this channel becomes insulating, and the source and drain are disconnected. We must note that when the input to the gate of the MOSFET is turned off, charge drains from the gate through the circuit it is connected to, and the MOSFET transistor turns off.

An FG-transistor is a MOSFET transistor that has no resistive connection to its gate. However, a control gate couples capacitively onto the transistor′s floating gate. In other words, an FG transistor has an additional electrode between the gate and the semiconductor. Unlike the other electrodes, however, the FG is not connected to anything. The cross section view of an FG-transistor is shown in Fig. 2.2. The channel current (i_d) of an FG transistor is a function of the voltage applied to the control gate (V_{cg}) and the amount of charge stored on the floating-gate node of the transistor. Schematic symbols of a typical MOSFET and an FG transistor are shown in Fig. 2.3 (a) and Fig. 2.3 (b), respectively. Fig. 2.4 shows the cross section view of the FG transistor including all parasitic capacitors. Assuming that no charge is stored in the FG:

$$Q = 0 = C_{FC}(V_{FG} - V_{CG}) + C_S(V_{FG} - V_S) + C_D(V_{FG} - V_D) + C_B(V_{FG} - V_B) \qquad (2.2)$$

where V_{FG}, V_{CG}, V_S, V_D, and V_B are the potential of floating gate, control gate, source, drain, and bulk respectively. We assume that $C_T = C_{FC} + C_D + C_S + C_B$ is the total capacitance of FG.

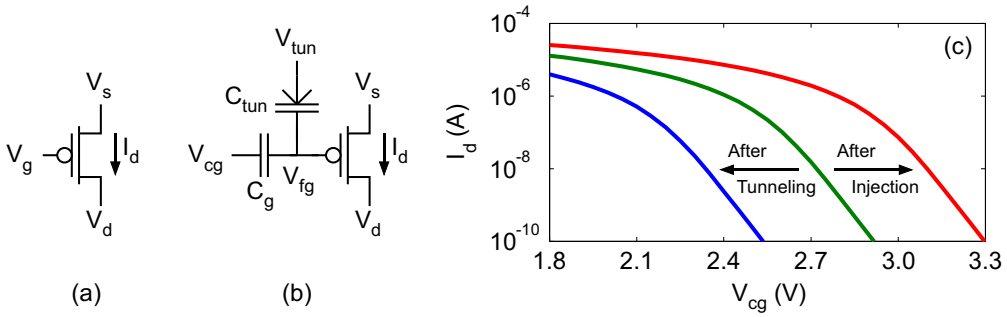


Figure 2.3: (a) Symbol of a typical transistor (b) Symbol of an FG transistor (c) Input-output characteristic of an FG-transistor

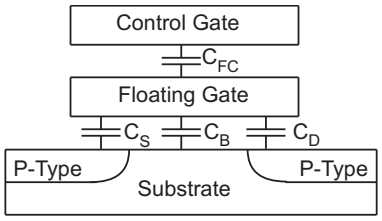


Figure 2.4: Cross section view of an FG-transistor with parasitic capacitors

We call $\alpha_j = C_J/C_T$, the coupling coefficient relative to the electrode J. The FG voltage can be expressed as:

$$V_{FG} = \alpha_G V_{GS} + \alpha_D V_{DS} + \alpha_S V_S + \alpha_B V_B \tag{2.3}$$

This equation shows that V_{FG} is a function of many node potentials. If the source and bulk are both grounded, V_{FG} can be expressed as:

$$V_{FG} = \alpha_G(V_{GS} + \frac{\alpha_D}{\alpha_G} V_{DS}) = \alpha_G(V_{GS} + fV_{DS}) \tag{2.4}$$

where

$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_{FC}} \tag{2.5}$$

We can translate device equations from a conventional transistor to an FG transistor. For example, the threshold voltage V_T and conductivity factor β, can be transformed to values measured with respect to the control gate.

$$V_T{}^{\mathrm{FG}} = V_T(Floating\,Gate) = \alpha_G V_T(Control\,Gate) = \alpha_G V_T{}^{\mathrm{CG}} \tag{2.6}$$

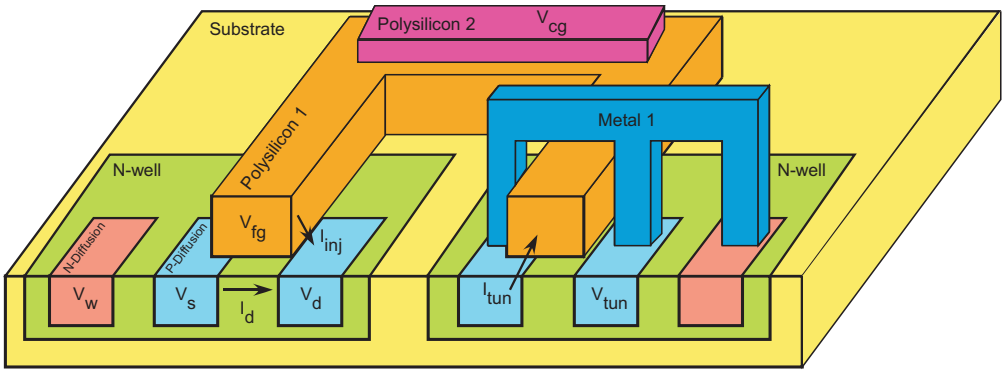


Figure 2.5: Cross section view of the FG-transistors used in this work

$$\beta^{\mathrm{FG}} = \beta(Floating\,Gate) = \frac{1}{\alpha_G}\beta^{\mathrm{CG}} \tag{2.7}$$

2.2.1 Structure of the FG transistors in this report

The cross-section view of the FG transistors used in our work is presented in Fig. 2.5. An FG transistor is a MOS gate surrounded by silicon-dioxide with no dc path to ground. The charge on the floating gate is stored permanently, providing a long-term memory cell since it is completely surrounded by a high-quality insulator. The control gate input of the FG transistor is created with polysilicon-2, and the floating gate of the FG transistor is created with the polysilicon-1. The actual FG transistor (M_{fg} in Fig. 2.6) is the left hand side pMOS in Fig. 2.5. Erasure and often writing is done by tunneling electrons through the tunneling junction, C_{tun}. The structure of this junction has a significant effect on the speed, efficiency, and long-term reliability of writing and erasure [10]. A thin oxide helps to reduce the tunneling voltage. In standard CMOS processes, the gate oxide of a CMOS transistor is preferred because it is thin and also of high quality, which benefits reliability and predictability. In Fig. 2.5 this tunneling junction is the right hand side transistor inside the substrate (M_{tun} in Fig. 2.6) with drain and source connected to each other. To remove electrons from the FG, V_{tun} is increased to a high voltage, usually higher than the reverse breakdown of the source/drain diffusions, but less than the breakdown voltage of the well-to-substrate junction. In order to avoid reverse breakdown, tunneling junctions are generally placed inside a well. The tunneling MOSCAP can have two basic types: a p^+ MOS capacitor formed as a standard pFET or a n^+ MOS capacitor formed with n^+ diffusion along the gate. In a 0.35μm CMOS process, we use pMOS transistors in an N-well to act as a tunneling junction because p^+ junctions are more consistent from process to process and the p^+ tunneling current is significantly higher [10].

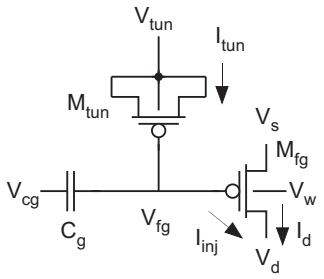


Figure 2.6: Schematic of the FG-transistors used in this work

2.2.2 Above threshold operation

The current to voltage relationship of the FG transistor in the above-threshold region can be expressed as follows. The condition to be in triode region is $|V_{DS}| < \alpha_G |V_{GS} + fV_{DS} - V_T|$, and the equation for drain current is:

$$I_{DS} = \beta[(V_{GS} - V_T)V_{DS} - (f - \frac{1}{2\alpha_G})V_{DS}{}^2] \tag{2.8}$$

The condition for the saturation region is $|V_{DS}| \geq \alpha_G |V_{GS} + fV_{DS} - V_T|$, and the drain current is:

$$I_{DS} = \frac{\beta}{2}\alpha_G(V_{GS} + fV_{DS} - V_T)^2 \tag{2.9}$$

where β and V_T of these two equations are measured with respect to control gate (β^{CG}, $V_T{}^{\mathrm{CG}}$).

There are some effects related due to the capacitive coupling effect, which modifies the FG I-V characteristics.

1. The FG transistor can go to the depletion mode even when $|V_{GS}| < |V_T|$. This is because of the effect of fV_{DS}. This effect is usually reffered to as "drain turn-out".

2. In conventional transistors I_{DS} is independent of the drain voltage in the saturation region. However, in FG transistors, I_{DS} will increase as the drain voltage increases and saturation does not happen.

3. The boundary between the triode and saturation regions for the FG transistor is expressed by

$$|V_{DS}| = \alpha_G |V_{GS} + fV_{DS} - V_T| \tag{2.10}$$

4. Unlike conventional transistors, g_m depends on V_{DS} in FG transistors.

$$g_m = \alpha_G \beta (V_{GS} + f V_{DS} - V_T) \tag{2.11}$$

5. The capacitive coupling factor depends on C_D and C_{FC} only. And its value can be verified by the following equation in the saturation region.

$$f = -\frac{\partial V_{GS}}{\partial V_{DS}} \tag{2.12}$$

2.2.3 Subthreshold operation

The following equation presents a quantified relationship for the channel current of the FG transistor in the subthreshold region.

$$I_d = I_0 \frac{W}{L} \exp\left(\frac{\kappa(Q_{fg} + C_g V_{cg} + C_d V_d + C_s V_s + C_{tun} V_{tun})}{C_{total} U_T}\right)\left[\exp\left(-\frac{V_s}{U_T}\right) - \exp\left(-\frac{V_d}{U_T}\right)\right] \tag{2.13}$$

where I_0 is the pre-exponential current scaler, κ is the subthreshold slope, U_T is the thermal voltage, C_g, C_d, C_s, C_{tun}, and C_w are the node capacitors, C_{total} is the total capacitance present on floating node, V_{fg}, V_{cg}, V_d, V_s, V_{tun}, and V_w are the node voltages, and finally, Q_{fg} is the charge stored on the floating node of the FG transistor. C_g is usually drawn such that it dominates C_{total} in order to make V_{cg} dominate the coupling term [10]. This will simplify V_{fg} to the following expression:

$$V_{fg} \approx \frac{Q_{fg} + C_g V_{cg}}{C_{total}} \tag{2.14}$$

Accordingly, the channel current expression will be simplified to the following expression:

$$I_d = I_0 \frac{W}{L} \exp\left(\frac{\kappa(Q_{fg} + C_g V_{cg})}{C_{total} U_T}\right)\left[\exp\left(-\frac{V_s}{U_T}\right) - \exp\left(-\frac{V_d}{U_T}\right)\right] \tag{2.15}$$

The V_d term is negligible, because the drain is typically connected to a low voltage. We can simplify the expression even more, resulting in:

$$I_d = I_0 \frac{W}{L} \exp\left(\frac{\kappa(Q_{fg} + C_g V_{cg})}{C_{total} U_T}\right) \exp\left(-\frac{V_s}{U_T}\right) \tag{2.16}$$

Therefore, a small change in the Q_{fg} results in some change in the effective threshold voltage of the FG transistor from the prospective of the control gate. This effect is illustrated in Fig. 2.3 (c).

This figure is a gate sweep on an FG transistor which is fabricated in $0.5\mu m$ CMOS process. In this experiment the FG transistor is programmed to three different values for Q_{fg}.

2.3 Reading Operation

Reading the stored charge (Q_{FG}) is done through applying a voltage to the control gate of the FG transistor and reading the current going through the channel. We will assume that there is some charge stored on the floating gate of the FG transistor. Therefore, we can say:

$$V_{FG} = \alpha_G V_{GS} + \alpha_D V_{DS} + \frac{Q}{C_T} \tag{2.17}$$

$$V_T{}^{\mathrm{CG}} = \frac{1}{\alpha_G} V_T{}^{\mathrm{FG}} - \frac{Q}{C_T \alpha_G} \tag{2.18}$$

$$I_{DS} = \beta[(V_{GS} - V_T - (1 - \frac{1}{\alpha_G})\frac{Q}{C_T})V_{DS} + (f - \frac{1}{2\alpha_G})V_{DS}{}^2] \tag{2.19}$$

Equation 2.19 shows how the injected charge can move the I-V curve of the FG transistor. Assume that the reading biases are fixed ($V_{GS} \simeq 5V$, $V_{DS} \simeq 1V$). Presence of the charge will greatly affect the level of the current that is sensed in the drain of the FG transistor. Fig. 2.7 shows two possible states that are used in Flash memories to demonstrate "0" or "1" cases. Curve A represents the "1" state and curve B is the same cell in the "0" state obtained with a 3-V threshold shift. In the defined reading condition I_D("1") is approximately $100\mu A$ and I_D("0") is around zero. Read operation is almost the same in analog applications. A fixed voltage is applied to the control-gate of the FG transistor and the current flowing through the channel of the FG transistor is measured using a measurement device.

2.4 Programming of FG Transistors

Many solutions are used to transfer electric charge (Q) from and into the FG. In both erasing and programming processes, we have to pass charge through a layer of insulating material. The charge stored on the floating node of the FG transistor can be modified using two famous phenomena, Fowler-Nordheim (FN) tunneling and hot-election injection. In FG transistors, these effects are used as efficient program/erase mechanisms.

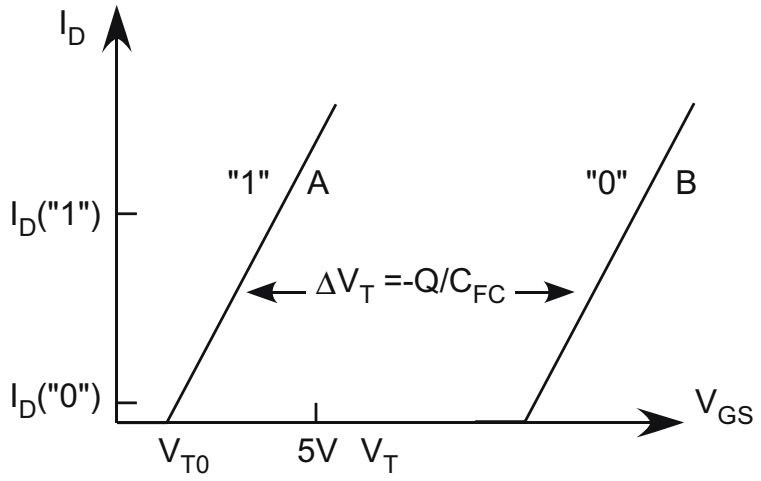


Figure 2.7: IV curves of an FG device when there is no charge stored in the FG (curve A) and when a negative charge Q is stored in the FG (curve B)

2.4.1 Tunneling Processes

In Fowler-Nordheim (FN) tunneling, a high positive voltage is applied to the tunneling junction of the transistor [11]. In this process, a large electric field is applied to electrons, allowing them to tunnel through an oxide [11]. The high voltage across the capacitor reduces the oxide thickness. With a relatively thick oxide (20 $\sim$ 30nm), we must apply a high voltage 20 $\sim$ 30V to have an appreciable tunnel current. With thin oxides, the same current can be obtained by applying a much lower voltage [7]. When the voltage across the tunneling junction is high, electrons are able to pass the barrier and tunnel through the oxide. C_{tun} shown in Fig. 2.3 (b) is the tunneling capacitor. The tunneling current is approximated by

$$J_{tun} \approx \alpha exp(-\beta t_{ox}/V_{ox}) \tag{2.20}$$

where t_{ox} is the oxide thickness, V_{ox} is the voltage across the oxide, and α=185.5A/μm^2 and β=32.8V/nm are fits that we have extracted across multiple processes and device sizes [10]. The tunneling capacitor that is used in this report is implemented using a simple MOS capacitor because the oxide in this capacitor is thinner than a typical poly-insulated-poly capacitor. This will reduce the voltage required to achieve FN tunneling. As an example, the voltages required for tunneling in 0.35μm is $V_{ox} > 8V$. Assuming that we have an array of FG transistors, to avoid write disturbs during the tunneling process, unselected FG transistors must either be disconnected from the tunneling input source using special high-voltage switches or the FGs of the unselected devices must be raised to a sufficient high-voltage that tunneling does not affect them. Since isolation of

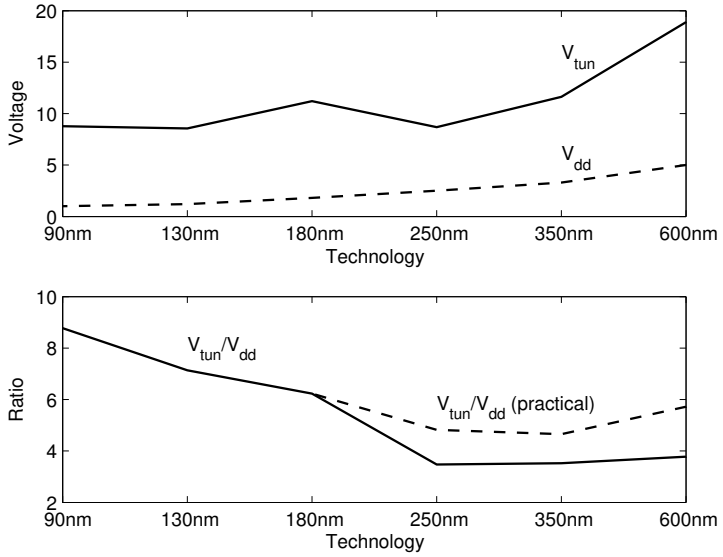


Figure 2.8: Scaling of write/erase voltages in standard CMOS. (a) Scaling of critical programming voltages: the Fowler-Nordheim tunneling voltage (V_{tun}) and the core supply voltage (V_{dd}). (b) Ratio of the tunneling voltage to the core supply voltage.

the FG transistors is a difficult task, tunneling is always a global process in analog memory arrays.

As (2.20) illustrates, the Fowler-Nordheim tunneling current depends on the oxide thickness, which decreases with CMOS process scaling, resulting in lower tunneling voltages in smaller technology nodes. However, charge retention in FG transistors is compromised by direct tunneling when $t_{ox} < 5$nm [12], which gives rise to the continued use of 5–7nm gate oxides in the newest NAND Flash processes while logic processes have continued scaling to a 2.5nm physical gate oxide thickness [13, 14]. Consequently, floating-gates that are built in standard processes below the 250nm node should use thick-oxide I/O devices that are rated for operation at 2.5V or greater so that oxide thicknesses are large enough to provide good charge retention.

Figure 2.8(a) shows the scaling of FG tunneling voltages in standard CMOS due to reducing the oxide thickness. The tunneling voltage (V_{tun}) was calculated for 1ms erase times using (2.14) and typical trends in oxide-thickness scaling (e.g. [15, 16]). Noting that the voltage required for generating a tunneling current is much higher than the rated supply voltage for each process, a step-up voltage converter, such as a charge pump, must be used. A charge pump typically multiplies the chip V_{dd} by a certain value, which is expressed by the step-up ratio (V_{tun}/V_{dd}). The necessary step-up ratios to achieve tunneling in each technology node are shown in Fig. 2.8(b). The step-up ratio is lowest for the 250nm through 600nm nodes because 1) below the 250nm node, higher-voltage I/O devices are used to make low-leakage floating gates and 2) above the 600nm node, devices are operated at 5V even though they can accommodate much higher voltages. However,

the rated V_{dd} of the 250nm through 600nm nodes produces noticeable hot-electron injection which can cause unwanted changes to the stored charge on an FG; consequently, these technology nodes must use lower supply voltages in practice, so the ratios for these nodes are essentially one integer value higher (see "V_{tun}/V_{dd} (practical)").

2.4.2 Injection Process

Channel hot-electron injection is used to insert electrons on the floating node of the FG transistors in many programmable analog systems [17]-[18]. Unlike the tunneling process, the injection process is only used to write to individual elements. Two main conditions must be met for injection to occur, a high current flowing through the transistor and a high electric-field between the gate and drain of the FG transistor. This high electric-field causes high-energy carriers to impact-ionize at the drain. A fraction of these ionized electrons distributes over the surface with high-enough energy to pass the oxide barrier and inject onto the floating gate.

As mentioned, to perform injection of an FG transistor, V_{dd} is increased to high voltages. All other voltages are also increased with respect to V_{dd}. This process is called "ramping up", in the remainder of this document. Obviously, the counterpart of this process will be referred to as "ramping down". [19] shows that the injection efficiency is better for subthreshold currents and it starts to drop in above V_T region. Hence for injection, the FG transistor is typically in the subthreshold region, and the injection current from the FG to the drain can be approximated as

$$I_{inj} \approx \beta I_s^{\alpha} e^{(}V_{sd}/V_{inj}) \tag{2.21}$$

where β, α, and V_{inj} are device-dependent fits [20]. Thus, programming speed is a function of V_{sd} and I_d, as well as V_{gd}, which is lumped into the fit parameters of (2.21). This current can also be expressed using the second equation as follows. The injection currents are functions of both the process parameters and the voltages applied to the FG transistor [11]-[21], as given by [22]:

$$I_{inj} = \gamma I_d \frac{V_{gd} + V_{TP}}{0.22 t_{ox}^{1/3} x_j^{1/2}} \exp\left(-\frac{\delta 0.22 t_{ox}^{1/3} x_j^{1/2}}{V_{gd} + V_{TP}}\right) \tag{2.22}$$

where $\gamma = 3$ and $\delta = 4.9 \times 10^8$ are fits that we have extracted across multiple processes and device sizes. Figure 2.9 shows how the voltages needed for injection ($V_{sd,inj}$) scale with technology nodes. Despite the stagnant oxide thickness, $V_{dd,prog}$ continues to scale because of X_j becomes smaller for

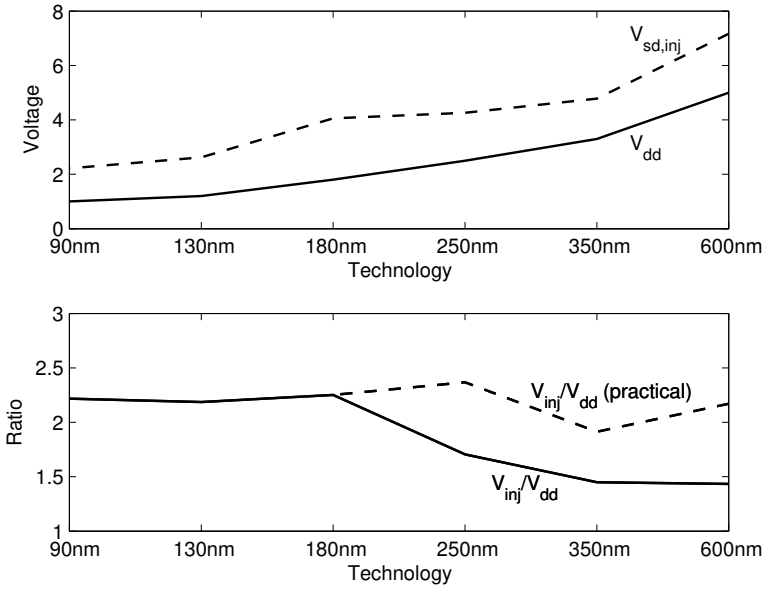


Figure 2.9: (Top) Scaling of critical programming voltages: the core supply voltage (V_{dd}), the write (i.e. injection) voltage ($V_{dd,prog}$), (Bottom) Ratio of the write ($V_{dd,prog}$) voltage to the core supply voltage.

smaller technologies.

Charge pump circuits are used to multiply the chip supply voltage by a constant and generate the voltage required for programming. A charge pump used for programming is enabled for a short period of time to minimize the energy consumption. The input to the charge pump is the chip supply voltage. The step-up ratio of the charge pump that is required for injection is a technology-dependent parameter and is shown in Fig. 2.9 (bottom). The step-up ratio is lowest for the 250nm through 600nm nodes because below the 250nm node, higher-voltage I/O devices are used to make low-leakage floating gates. However, the rated V_{dd} of a technology node causes too much hot-electron injection in standard thin-oxide transistors for stable floating-gate operation. Consequently, the 250nm through 600nm nodes must use lower supply voltages in practice, so the ratios for these nodes are essentially one integer value higher than shown in the figure. Since the programming infrastructure [2] requires an additional overhead of several hundred millivolts beyond the necessary V_{sd} for injection, a charge pump for injection must be able to provide $(2 \sim 3) \times V_{dd}$. In addition, in order to have a suitable charge pump for the injection process, we must have a circuit that has low output ripple. Some other aspects of a good charge pump are fast start-up time and small die area. A charge pump proposed for injection will be presented in the next Chapter.

2.4.3 Injection Methodologies

There is no standard way of performing FG injection. However, there are two major injection techniques that some research groups employ: Pulsed-Programming and Continuous-Programming.

Pulsed-Programming

In pulsed-programming, a short programming pulse is applied first and then the programmed charge is measured in the next phase. Fig. 2.10-(a) shows the pulsed-programming technique. This process continues until the target current is reached. This approach requires a lot of programming circuit overhead for reading and programming modes. Also, it takes longer to program higher target currents. Therefore, programming time is a function of target current. This will be a serious issue when high programming accuracy is required. Because in each programming pulse, a finer amount of charge must be added to the FG node, which will increase the number of programming pulses. This limitation will be more problematic when a large array of FG transistors must be programmed. In order to mitigate this issue, a few approaches have been presented before [23, 20, 19]. For example, the approach presented in [19] was tried in $0.25\mu m$ and $0.5\mu m$ n-well CMOS processes and can be used to program a large FG array with more than 0.2% of accuracy over a wide range of currents (more than 3.5 decades). In this approach, programming time is minimized by choosing an optimal injection rate for a specific target current. Thus, a characterization of the injection rate is done first in calibration mode, and then the developed mathematical model summarizes the characterization data into a few parameters which simplify the process. Accordingly, run-to-run and chip-to-chip mismatch of different injection rates for different FG transistors are taken into account.

The pulsed-programming proceeds as follows. The initial current flowing through each target FG transistor is measured first. Then, for each initial current and the target current, the optimal V_{ds} is calculated using the data produced in the calibration process. Next, the chip is ramped up and each target FG transistor is programmed using the calculated V_{ds} voltage. Then, the chip is ramped down and the currents are measured. If the measured current is less than the expected value, the previous steps will be repeated until each FG transistor is injected to the desired value. In this scheme, the target current is reached in a maximum of ten steps [19]. The required V_{ds} for each element for each step of the injection process is calculated using the data achieved in the calibration process. [19] shows that the programming process was successful for both deep subthreshold and above threshold target currents.

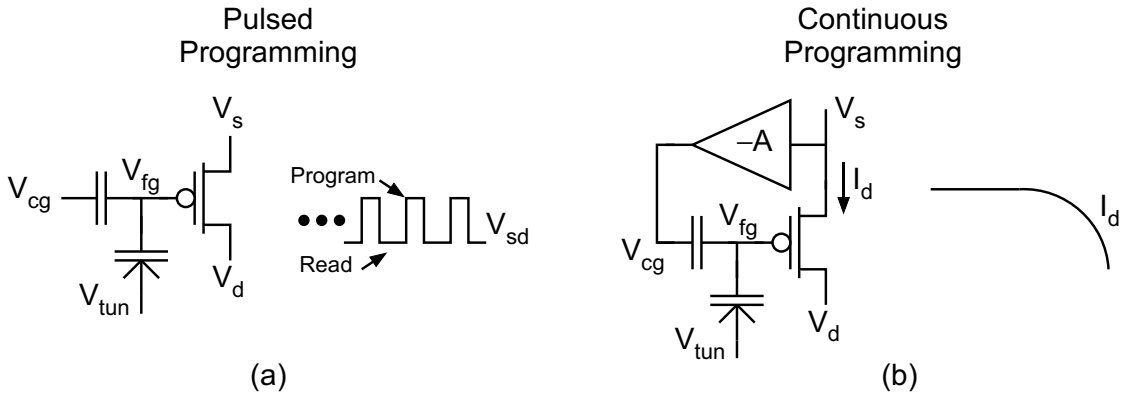


Figure 2.10: Comparison of pulsed and continuous programming techniques. (a) Pulsed programming (b) Continuous-time programming.

All the pulsed-based programming techniques presented in [23, 20, 19], will reduce the programming time; however, they still suffer from some drawbacks. For example, in [23] a high accuracy voltage source is required (> 12-bit precision). In addition, [19] requires a calibration phase to extract a few parameters that are required in the programming phase to choose the correct V_{sd} for programming pulses. This makes the programming process complicated and makes it difficult to be used in a large array of FG transistors. In addition, this technique requires high-precision data converters, and wide-range current measurements, which complicate the system and make it not suitable for portable applications. One advantage of pulsed-programming is that the FG is measured in a state similar to run-mode; with no high program voltages applied to the FG cell and with the same current levels that will be used in run mode. This will improve the accuracy of the programming.

Continuous Programming

Continuous methods of programming are faster than pulsed-programming techniques and require less peripheral circuits to program an entire array of FG transistors. In continuous methods of programming, a high voltage is applied to the FG transistor continuously and a negative feedback is used to make the programmed voltage converge to the target voltage. In this technique, the FG transistor is usually used in a source-follower configuration and the injection is linearized via negative feedback to the control gate. Fig. 2.10-(b) shows an example of a continuous programming method. Fig. 2.11 shows the schematic of the amplifier that is used in the negative feedback path of [1]. In [2], the same characteristics are achieved using a single transistor in the feedback path as shown in 2.12. This structure is smaller and gives more flexible control over the injection rate

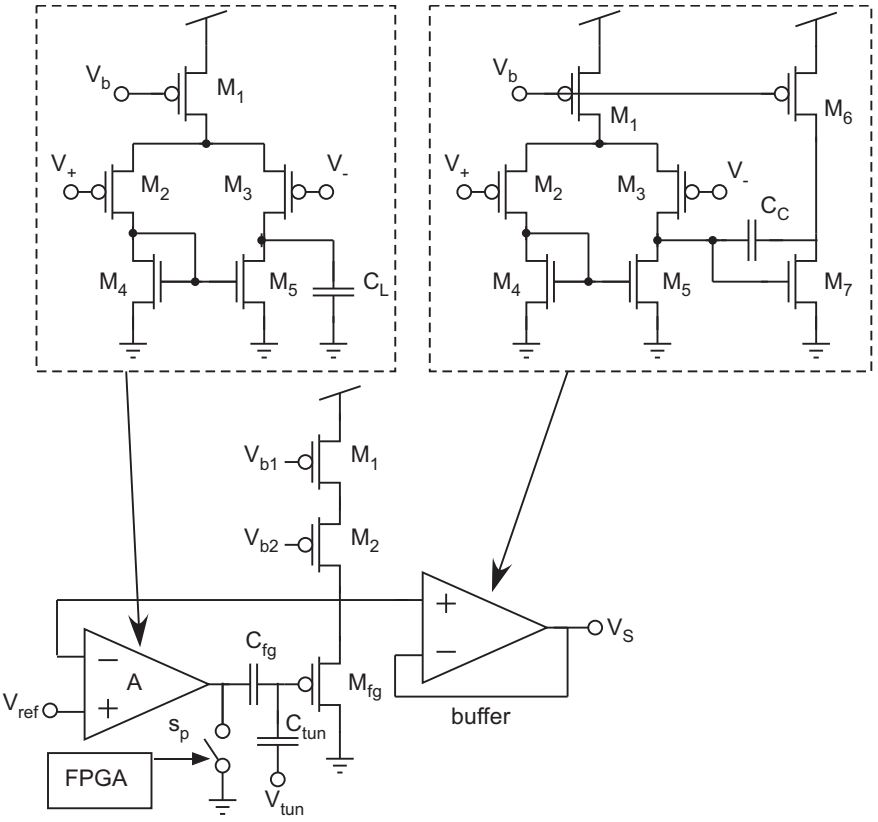


Figure 2.11: Schematic of the FG transistor structure used in [1] to linearize the programming process

since V_s can be modified using either a voltage or a current input.

There are a variety of continuous-time programming structures available in the literature. A single-transistor circuit [20] is an example, which self-converges due to the negative feedback of injection current from the FG to the drain. The memory cell presented in Fig. 2.12 does not converge on its own, and it needs some external programming circuit. In [2] a feedback structure is used to stop the continuous-time FG programming.

2.5 Standard Flash Memories

In general, Flash memories used for digital applications are subdivided into two major categories: NAND and NOR. Based on how the FG transistors are organized in an array, it is possible to distinguish between NAND flash memories and NOR flash memories. NOR flash emerged in 1978 by Intel and changed the market which was dominated by EPROM and EEPROM devices. NAND Flash architecture was presented by Toshiba in 1984. Many traditional embedded systems have used NOR flash for Non-volatile memory. Most of the current designs are moving to NAND flash from NOR flash, because of its higher density and lower price for high-performance applica-

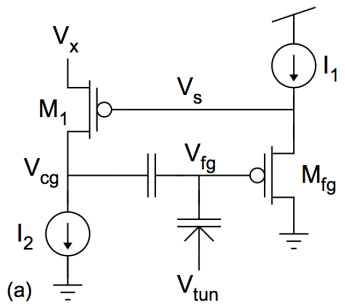


Figure 2.12: The floating Gate memory cell used in [2]

tions. In the next few subsections, I try to cover the differences between NAND and NOR flash memories and pros and cons of each memory type.

2.5.1 NAND Flash Memories

In the NAND flash architecture, the FG transistors are connected in a series combination, in a group of 16 or 32. Two selection transistors are located at each end of the string, to ensure connection to ground through M_{SSL} and to the bitline through M_{DSL}. Fig. 2.13 shows a basic structure of the NAND memory cell. While reading a cell, its gate voltage must be at 0V, and other gates of the string are biased with a high voltage (usually 4-5V). In this way, they work as a pass-transistor, regardless of their threshold voltage. The threshold voltage of an erased NAND Flash Cell is a negative voltage. A programmed cell has a positive threshold voltage, which is typically less than 4V. Basically, if the addressed cell is erased, driving the selected gate with 0V causes a sink current in the transistor. On the other hand, if the addressed cell is programmed, no current is sunk [24].

The NAND Flash array can be grouped into a series of blocks, which are the smallest erasable cells is NAND Flash devices. When erasing a block, all bits become 1 and all bytes become hex "FF". By programming, each bit changes from 1 to 0. The smallest programmable cell is a byte. Usually, each NAND block can survive 100,000 programming/erasing cycles. NOR flash has 10 times less life span.

2.5.2 NOR Flash Memories

A NOR flash memory is basically an FG transistor, programmed by channel hot-electron (CHE) injection and erased by Fowler-Nordheim tunneling. In this type of memory cell, the neutral (or positively charge) state resembles logical state "1" and the negatively charged state with the

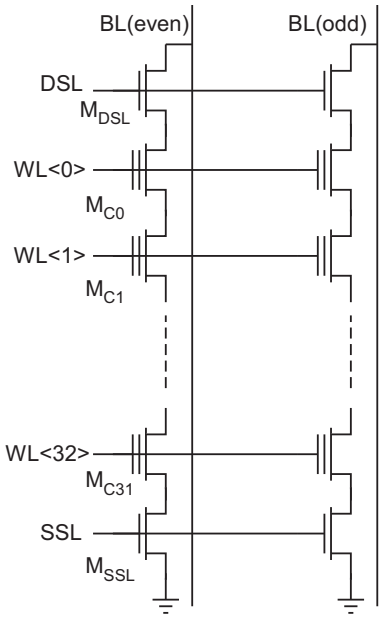


Figure 2.13: Matrix structure in NAND architecture

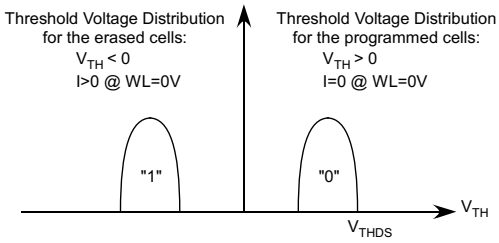


Figure 2.14: Threshold voltage distribution for erased and programmed cells

negative charge stored in the FG, resembles "0" state. The name "NOR" comes from the structure that is used to arrange the cells inside an array, through rows and columns in a NOR-like structure. The gate connections shared between the Flash cells is called a wordline (WL), while the drain terminals shared between Flash cells form the bitline (BL). In this structure, the source terminal is common to all of the cells. A NOR Flash array is shown in Fig. 2.15. Programming of the NOR memory cells is done through channel hot electron (CHE) injection in the FG at the drain side. This process for a single cell is shown in Fig. 2.16. In this process, electrons gain high energy to pass the oxide-silicon energy barrier, due to the electric field in the channel between the source and drain terminals. The Fowler-Nordheim electron mechanism is a quantum mechanical tunnel produced by an electric field. Applying a strong electric field (around $8 \sim 10MV/cm$) across a thin oxide, it is possible to force a large electron tunneling current through it without destroying its dielectric properties. This process is used to remove electrons from the FG to silicon surface as shown in Fig. 2.17.

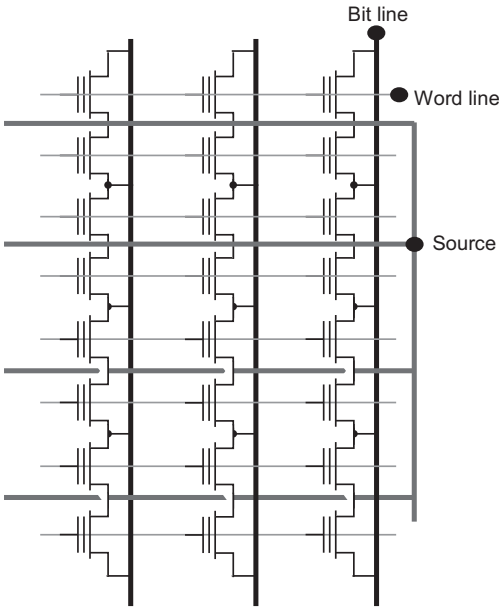


Figure 2.15: NOR Flash array equivalent circuit.

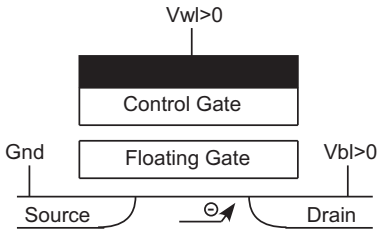


Figure 2.16: NOR Flash programming process.

2.5.3 NAND Flash vs. NOR Flash

Fig. 2.18 shows a brief comparison between NAND flash and NOR flash cells. The size of a NAND flash cell ($4F^2$) is smaller than a NOR flash cell ($10F^2$) because a NOR flash needs a separate metal contact for every single cell. This will provide higher densities for low-cost consumer devices in a significantly reduced die area i.e. removable cards, including USB drives, secure digital cards (SD), memory stick cards, and multimedia cards (MMCs). However, NOR flash memories are typically used for code storage and execution, like simple consumer appliances, and low-end cell phones, and embedded applications.

Basically, NOR flash is a random access memory device, because it has enough address pins to map its entire media which allows for easy access to every single bit of its bytes. NAND devices are interfaced serially via a complicated I/O interface, which varies from one device to another. Also, the access time of the NAND flash memories is much higher than the NOR flash memories. The reason behind the slower access time of NAND flash memories is that the sense amplifier of a

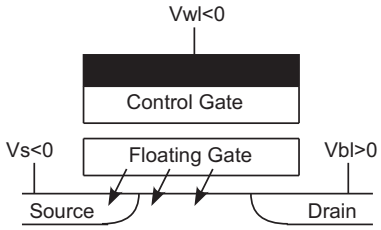


Figure 2.17: NOR Flash erasing process.

Table 2.1: NAND Flash vs. NOR Flash

	NAND	NOR
Advantages	Fast Program	Random Access
	Fast Erase	Byte Programs possible
Disadvantages	Slow random access	Slow Programs
	Byte Programs Difficult	Slow Erase
Applications	File (disk) applications	Replacement of EPROM
	Voice, data, video recorder	Execute directly from Non-Volatile Memory
	Any large sequential data	

NAND configuration sees a signal weaker than for a NOR configuration. This is because multiple transistors are in series. This weak signal slows down the speed of the reading process. This can be overcome by operating in serial access mode. NAND flash is suitable for a file or sequential data applications. NOR flash is appropriate for random access applications. Random access time on NOR flash is around $0.075\mu s$ and on NAND flash (for the first byte only) is $25\mu s$ [25]. However, after initial access has been made, the rest of 2111 bytes are shifted out of memory at speed of $0.025\mu s$ per byte. NAND flash has faster "Programming" and "Erasing" operations. Random-Access and Byte-write capabilities are the main advantages of the NOR flash memories [25]. Table 2.1 summarizes the pros and cons and applications of both types.

2.6 Analog Applications of FG transistors

Other than Flash-memories, which is a digital application of FG transistors, FG transistors have been used in a wide variety of analog applications, too. For example, [26] is a programmable analog filter bank that is realized with FG transistors. Analog filter banks are used in audio and vibration-sensing applications. In these applications, frequency analysis must be performed. The system that performs this analysis must be low power and have moderate to high precision. FG transistors are used as bias current circuits in [26]. [18] also presents a field-programmable-analog-array (FPAA) for wireless sensor hardware applications. In [18], FG transistors are also used to provide accurate bias currents for the circuits with programmable parameters. These parameters

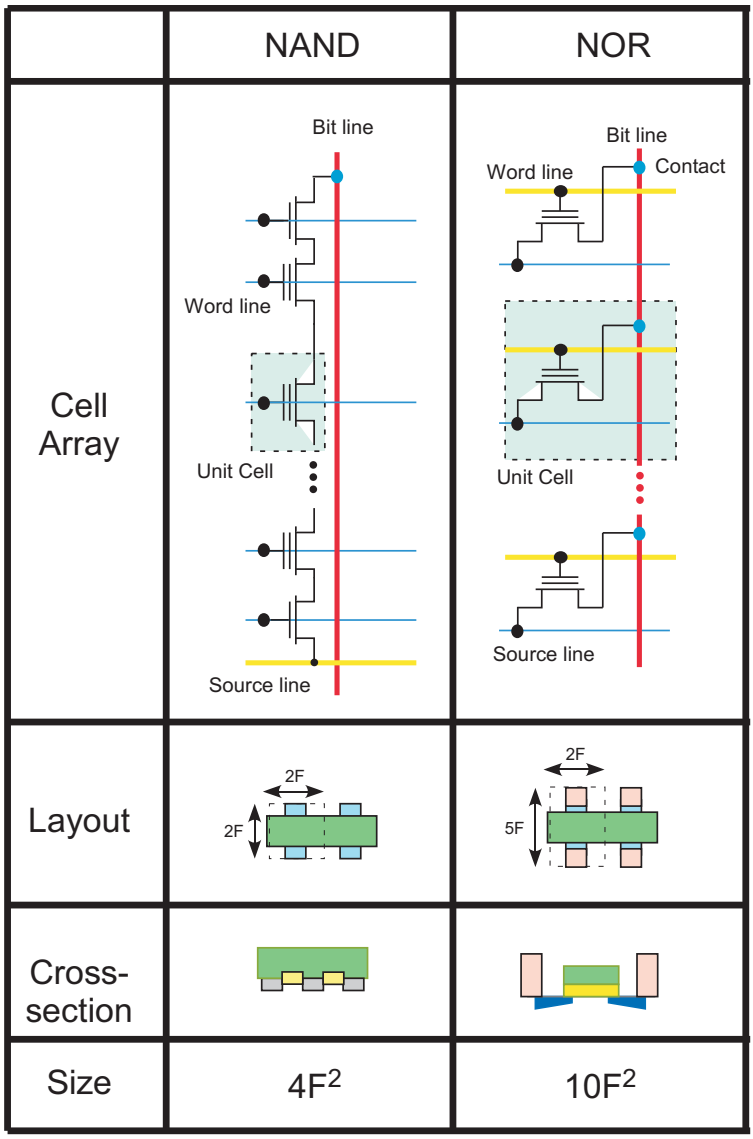


Figure 2.18: Comparison of NAND and NOR Flash Cells

include the gain of OTAs, the bandwidth of a filter, etc.

[27] uses FG transistors as analog memory cells to monitor the bone healing process. In this work, the possibility of self-powered monitoring of the bone healing process using piezo-floating-gate (PFG) sensor is investigated. The data that was recorded by the sensors includes the statistics of the strain evolution during the healing process and is stored in an array of FG transistors for offline retrieval and analysis. [28] also uses FG transistors as analog storage. In this work, the data related to the estimated center of mass (centroid mean) and spread (centroid variance) are stored in the FG transistors.

FG transistors are also used in analog to digital converters (ADCs), too. The resistive divider used in flash ADCs have some limitations. The mismatch between different resistors in the resistive

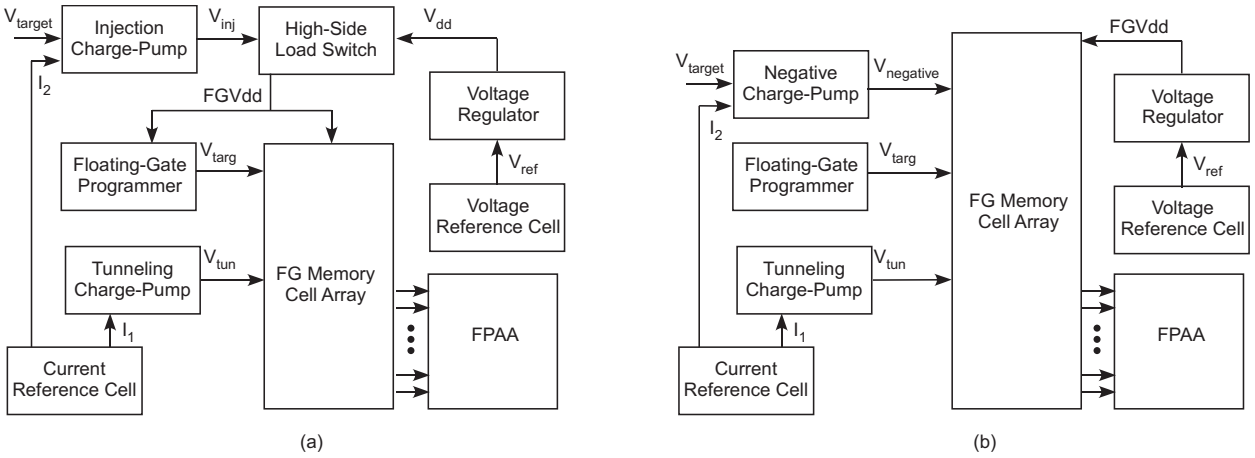


Figure 2.19: Two possible reconfigurable systems using FG transistors.

divider results in incorrect and non-linear voltage division. [29] presents an array of adaptive FG comparators. In this structure, each reference voltage is stored and programmed separately using FG technology. The quantizer made of this structure is called the Adaptive FG Quantizer (AFGQ). The same research group presents an adaptive FG comparator with 13-bit resolution in a different work [30]. The comparator presented in [30] uses non-volatile charge storage for either offset nulling or automatic programming of the desired offset voltage. The negative feedback functionality of pFET hot-electron injection is used to achieve fully automatic offset cancellation.

2.7 Objective of the proposed research

The main objective of this work is to use FG transistors and make a low-power reconfigurable system for portable applications. An FPAA is used to make a reconfigurable system. An FPAA is an integrated circuit including configurable blocks (CAB) and interconnects between these blocks. The circuits used inside the CABs can be either voltage mode or current mode devices. The bias current required for these devices are provided with FG transistors. The reason that we do not use on-chip current sources instead of FG transistors is that the FG transistors provide more flexibility in the bias current generation. Some infrastructure circuits are required to program the FG transistor. These circuits include charge pump circuits to generate high voltages required for tunneling and injection, a voltage reference cell to generate the bias voltage required for the rest of the circuits, voltage regulators to generate low noise supply voltages with high current driving capability, and the programming circuits to control the injection rate and accuracy of programming the FG transistors. Two possible block diagram of the programming infrastructure are shown in

Fig. 2.19. In Fig. 2.19(a) a positive charge pump is used to perform the injection of the FG transistors. This charge pump is replaced with a below ground charge pump in Fig. 2.19(b) to improve the accuracy of programming. In the next few Chapters the design procedure of most of these blocks used in Fig. 2.19 will be covered.

Chapter 3

High Voltage Charge Pumps for Tunneling and Injection of Floating-Gate Transistors

Tunneling and injection of FG transistors require high voltages. The high voltage required for a specific programming process can be generated on-chip or off-chip. A high voltage can be generated using off-chip boost converters [18]. In this chapter, we discuss how we generate the high voltages required for programming using high voltage charge pumps. These charge pumps do not require bulky on-chip inductors. In the rest of this chapter, three on-chip charge pumps are presented, which are designed in 0.35μm CMOS technology for tunneling and injection, respectively.

3.1 A Regulated Charge Pump for Tunneling Floating-Gate Transistors

In this section, we present a charge-pump topology fabricated in a standard 0.35μm CMOS process that is capable of providing the high voltages required for programming floating-gate transistors. We focus primarily on a charge pump for the tunneling operation since tunneling requires a high step-up ratio. Because the voltage across any single device in this charge pump never exceeds V_{dd}, this topology can easily be scaled for higher or lower output voltages, hence providing a means to produce the voltages required by other high-voltage on-chip applications, such as injection of floating-gate transistors or for MEMS devices. We also provide insight into the voltages that are required to generate tunneling conditions in floating-gate devices fabricated in standard CMOS processes as these processes scale to newer technology nodes.

Additionally, this charge pump has been designed for use in battery-powered applications in which power consumption must be minimized and the charge pump should be robust to noisy and/or drooping battery supplies. To meet these needs, we describe several circuit techniques, including 1) a simple method for reducing start-up energy in a previously reported charge-transfer stage, 2) a method for regulating the charge pump via frequency modulation, and 3) a new circuit to minimize the short-circuit current of the clock circuitry. The topology we present can be easily designed to achieve a specific, and low, output ripple that is independent of the clock frequency and load current—low output ripple is crucial to precisely programming floating-gate transistors for applications in non-volatile analog memory and programmable analog biases. The resulting charge pump is able to provide a low-ripple output up to 16V from a 2.5V supply while expending only $1.45\mu J$ to erase floating-gate transistors.

3.1.1 Considerations Regarding Tunneling Voltages

Fowler-Nordheim tunneling is the process of moving electrons through an oxide using a large electric field [11]. This process is often used in floating-gate transistors to modify the charge stored on the floating gate and, accordingly, change the memory value. In many floating-gate applications in standard CMOS processes, tunneling is used for memory erasure [31, 19].

As we discussed in the previous Chapter, Fig. 2.8 provides insight into the voltages that need to be generated by the charge pump for the tunneling voltage. The amount of load current required to tunnel a single FG transistor is exponentially related to the inverse of the voltage across the tunneling oxide. Since the starting charge on the floating gate may be unknown and different for each tunneling operation, the current required to tunnel a device can vary greatly for each tunneling operation. Additionally, since tunneling is often used as a global erasure [19], many floating-gate transistors will likely be tunneled simultaneously by the same charge pump. With each floating-gate charge starting at different values, the load current of the charge pump cannot be known ahead of time and can vary by several orders of magnitude. Consequently, the charge pump providing the tunneling voltage must be able to provide a consistent voltage for a wide range of load currents.

Additional constraints on the charge pump that are specific to embedded analog applications include small size with minimal external components, reasonably fast start-up, and shut-down, and consistent voltage generation in the presence of potentially noisy supply voltages; voltage consistency is especially important since it affects the programming accuracy. For generating

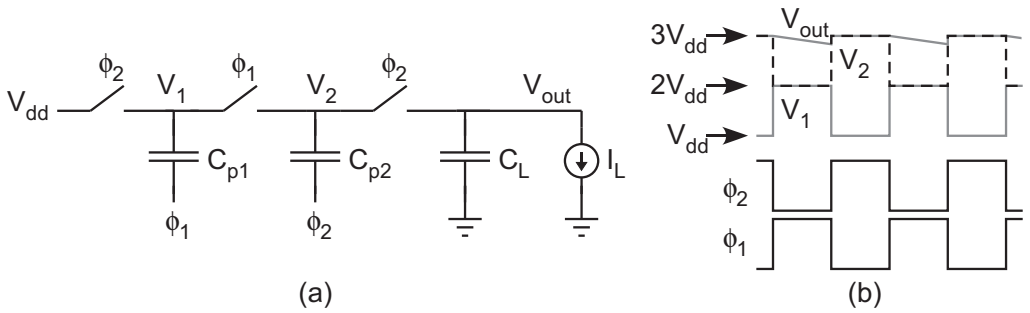


Figure 3.1: (a) Ideal charge pump. (b) Operation of the ideal charge pump.

tunneling voltages, which exceed the breakdown voltage of the devices, care must be taken in the design of the charge pump so that no individual device undergoes breakdown. Furthermore, for battery-powered applications, energy consumption should be minimized; therefore, the charge pump should only be enabled when it is needed to provide the high tunneling voltage. Finally, this charge-pump technology should be scalable to the tunneling-voltage requirements of other technology nodes. We demonstrate a charge pump in a standard 0.35μm CMOS process because of our existing applications of programmable and reconfigurable analog systems in this process [32, 18], and this same topology can be scaled to meet the needs of tunneling voltages in other processes, as well. In the remainder of this section, we describe our design of an integrated high-voltage charge pump that is capable of meeting all these requirements.

3.1.2 Brief Overview of Charge Pumps

A charge pump, which is sometimes referred to as a voltage multiplier, is a switched-capacitor voltage converter that is often used to create voltages outside of the typical voltage rails (e.g. above V_{dd}). Figure 3.1 shows an idealized Dickson charge pump[33], which is a standard charge-pump topology for integrated circuits because of its linear voltage growth and its insensitivity to stray bottom-plate capacitance. This charge pump has two stages that are clocked by alternating clock phases (ϕ_1 and ϕ_2). When ϕ_1 is low (ϕ_2 is high), node V_1 and capacitor C_{p1} are charged to V_{dd}. On the subsequent clock phase, ϕ_1 transitions high, which raises the bottom terminal of C_{p1} to V_{dd}, thereby "pumping" V_1 to $2V_{dd}$; this resulting voltage at node V_1 is sampled at node V_2. When ϕ_2 goes high again, node V_2 is raised to $3V_{dd}$. The final output voltage is obtained by sampling the last stage onto the load capacitor, C_L. Higher voltages can be generated by cascading more stages, N. Each stage adds V_{dd} to obtain a total output voltage of $V_{out} = (N + 1)V_{dd}$.

However, when a load current, I_L, is drawn from the charge pump, V_{out} is reduced. In equilib-

rium, the load draws a charge of $I_L T$ during each cycle of duration T which results in a voltage loss of $I_L T/C_p$ for each pumping capacitor. The resulting output voltage is

$$V_{out} = (N+1)V_{dd} - N\frac{I_L}{C_p f} \tag{3.1}$$

where $f = 1/T$ is the pumping frequency[34].

3.1.3 Charge-Transfer Switches

The primary challenge when designing a charge pump that approaches the ideal characteristics in (3.1) is the design of the charge-transfer switches (CTS), which are illustrated as ideal switches in Fig. 3.1. Early integrated charge pump designs used diodes or diode-connected transistors to implement the CTS[33]. Such designs rely upon the uni-directional current flow of diodes to only allow charge transfer onto a pumping capacitor when its voltage is exceeded by the voltage of the preceding stage. These designs suffer from poor voltage gain and poor efficiency because of the accumulation of diode voltage drops. As a result, several charge-pump circuits have been developed to dynamically control the on/off state of the CTS and, thus, improve performance.

As described in section 3.1.1, Fowler-Nordheim tunneling requires voltages so large that the drain-to-body junction will break down. To avoid break down, transistors in the CTS must inhabit isolated wells so that all voltage differentials in the CTS are at safe values. Since it is not always possible to isolate nFETs in standard CMOS processes, it is best for high-voltage charge pumps to use pFETs exclusively.

The all-pFET CTS that we use is based upon the circuits presented in [35, 36], and we have modified them to improve performance—particularly in reducing start-up power consumption. Our resulting charge-pump stage is shown in Fig. 3.2(a). Each stage has two parallel paths—a top path through M_{sw1} and M_{sw3}, and a bottom path through M_{sw2} and M_{sw4}. The parallel paths conduct in opposite phases, which helps to reduce the output ripple. Furthermore, the opposing phases of the parallel paths offer a low-complexity means for clocking the second set of switches (M_{sw3} and M_{sw4}). This second set of switches reduces the voltage stress on the transistors in the off-phase. In a CTS with a single series switch, the voltage across an off switch is $2V_{dd}$. By adding the extra series switches, the off voltage is divided across the series switches so that no pair of terminals on the transistors is exposed to a voltage higher than V_{dd}[35, 36].

To provide the correct voltage for the transistor wells, the active well-biasing technique is used

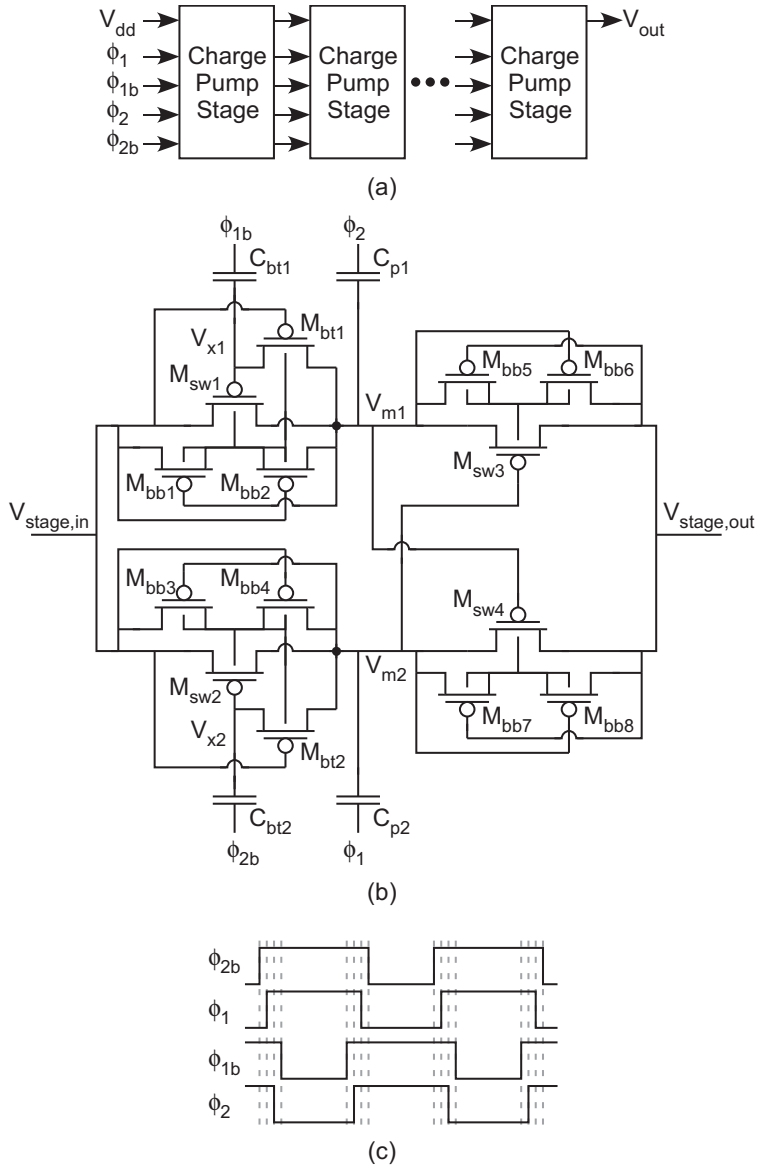


Figure 3.2: (a) The open-loop charge pump consists of a cascade of charge-pump stages. (b) The all-pFET charge pump stage that is used throughout this work. (c) Non-overlapping clock signals necessary for the charge pump stages.

[37]. This technique is implemented by the "bulk-biasing" transistors M_{bb*}. Each pair of M_{bb*} transistors connects the well to the higher voltage terminal of the source or the drain. The system that was simulated in [36] extended the CTS of [35] by including active well biasing for switches M_{sw1} and M_{sw2} but not for M_{sw3} and M_{sw4}; instead, the wells of M_{sw3} and M_{sw4} were connected to $V_{stage,out}$. In steady-state, this was acceptable because they included a grounded capacitor at $V_{stage,out}$ to hold the higher of V_{m1} and V_{m2}. However, we have removed this capacitor and added well-biasing on these switches to avoid charging extra capacitance and activating parasitic vertical BJTs during startup, both of which unnecessarily consume power. Reducing the startup power is

Table 3.1: Charge Pump Specifications

Technology	0.35μm CMOS
V_{dd}	2.5V
# Stages	6
C_{bt}	110fF
C_p	1.5pF
M_{bb}	3μm x 0.35μm
M_{bt}	3μm x 0.35μm
M_{sw}	5μm x 0.35μm

important for charge pumps that are used for tunneling floating-gate transistors; startup energy dominates the overall energy of the charge pump, particularly since the charge pump is turned on for relatively short periods of time while tunneling.

The operation of this charge-pump stage is as follows, where the phases of the 4-phase non-overlapping clock are depicted in Fig. 3.2(c). Focusing only on the top half of the circuit and noting that the bottom half is identical (and out of phase), we can see that charge is transferred from $V_{stage,in}$ to V_{m1} when M_{sw1} is turned on (ϕ_{1b} is low). When ϕ_2 goes high, node V_{m1} is increased by V_{dd}. This charge is then transferred to $V_{stage,out}$ when M_{sw3} is on (ϕ_1 is low). M_{bt1} is used to ensure that the correct voltage is at the gate of M_{sw1} at all times so that M_{sw1} is either fully on or fully off.

3.1.4 Complete Open-Loop Charge Pump

The complete, open-loop charge pump consists of six identical stages in series [Fig. 3.2(a) Top]. Although each charge-pump stage contains many devices, its size is dominated by the pumping capacitors C_{p*}. Each C_p is only half of the value needed for a charge pump with one capacitor per stage[34], so the size is similar to other charge pumps with commensurate performance. The design specifications are summarized in Table 3.1.

Figure 3.3(a) shows the output voltage of the six-stage open-loop charge pump for varying clock frequencies and load currents, and Fig. 3.3(b) shows the output resistance (which will be discussed in more detail in section 3.1.7). As can be seen from Fig. 3.3(a), this charge pump is able to achieve the necessary voltages for tunneling a floating-gate transistor (approximately 10V-12.5V in this 0.35μm CMOS process).

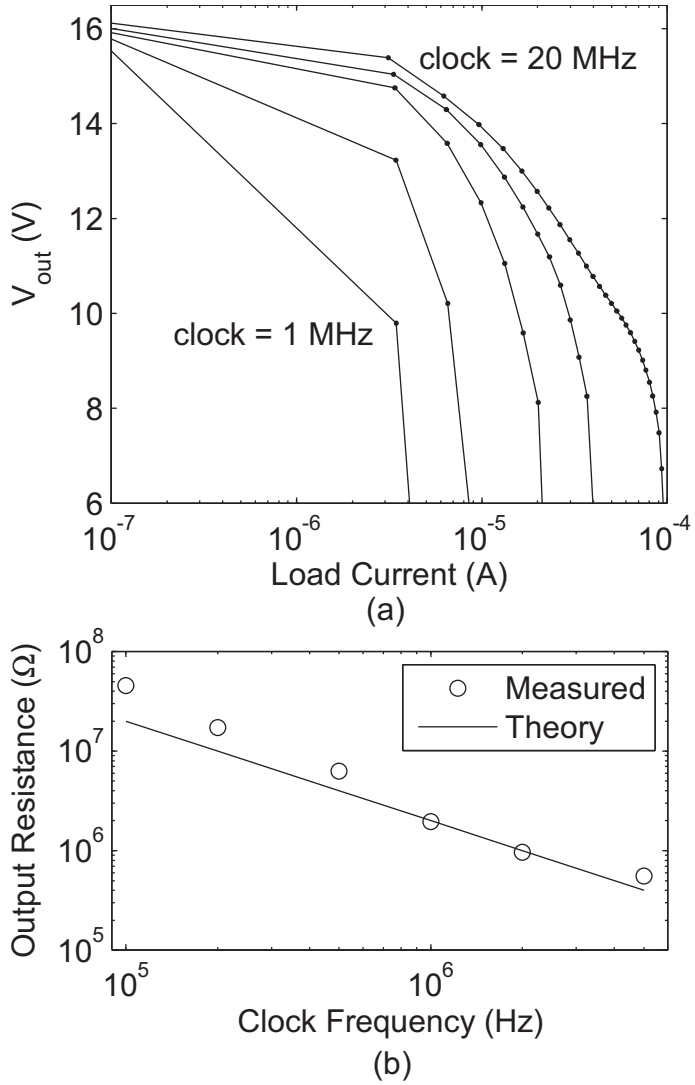


Figure 3.3: Measured characteristics of the open-loop charge pump. (a) Output voltage characteristics as a function of load current for multiple clock frequencies ∈ [1MHz, 2MHz, 5MHz, 10MHz, 20MHz]. (b) Output impedance of our charge pump as a function of clock frequency.

3.1.5 Regulation Through Frequency Variation

When a charge pump is used to generate tunneling voltages, V_{out} must be steady and consistent to facilitate accurate floating-gate programming. However, referring back to the idealized charge pump of Fig. 3.1, it is evident from (3.1) that V_{out} has an amplified dependence on the supply voltage, V_{dd}, which may be inconsistent and noisy in a battery-powered application. V_{out} also has a dependence on the load current I_L, which will vary as memory cells begin and finish programming; this dependence on the load current is clearly evident from the results of the open-loop charge pump [Fig. 3.3(a)] and can significantly vary the output voltage. Furthermore, in an open-loop charge pump, V_{out} is set to integer multiples of V_{dd}, which is a limitation for setting the sensitive program voltages to optimal values. To achieve reliable and accurate tunneling programming, the

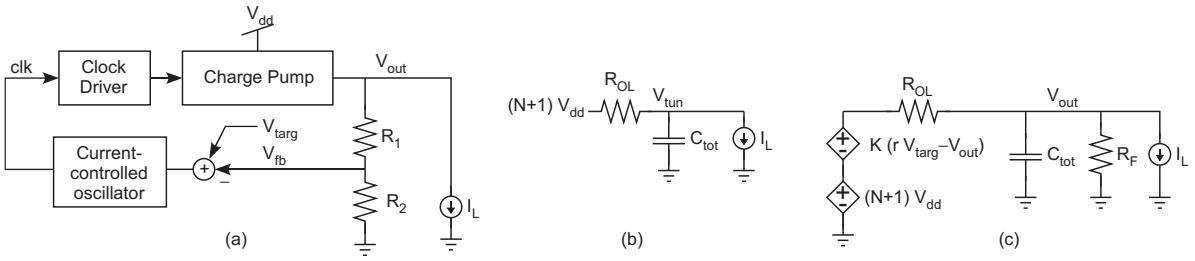


Figure 3.4: (a) Block diagram of a regulated charge pump. (b) Linear model of an unregulated charge pump. (c) Small-signal model of a regulated charge pump.

charge pump should be regulated to a constant output voltage.

Examining (3.1), only two quantities can be adjusted at runtime to regulate the output voltage: V_{dd} and f. Regulation using V_{dd} is typically accomplished by modulating the clocking voltages ϕ_1 and ϕ_2[38, 39]. These clocking voltages contribute the NV_{dd} term of V_{out} in (3.1). These variable-pump-voltage regulators have the advantage of reducing the level of clock-feedthrough ripple on V_{out}, which must otherwise be removed with a large load capacitor. However, variable-pump-voltage regulators have the disadvantage that they constantly operate at their maximum frequency, which results in wasted power from unnecessarily charging and discharging all parasitic capacitances.

Variable-frequency regulators are thus a more efficient alternative. The simplest type of variable-frequency regulator is the "skip" regulator, which turns on a constant-frequency oscillator when V_{out} is less than the desired voltage and otherwise turns the oscillator off[40]. Some regulators have used a combination of variable-voltage and skip-mode [41, 42]. Skip regulators exhibit sporadic bursts of pumping which create a large ripple on the output. A better alternative is a true "variable-frequency" regulator that linearly increases or decreases the frequency to regulate V_{out}. This form of regulation has been previously used in [43, 44], and we have used the same basic method. In contrast to the voltage-doubler charge pump of [43] that used very-large external capacitors, we designed for small size and efficient operation at the high voltages/low load currents typical of tunneling. Furthermore, our charge pump achieves higher efficiency, smaller size, and better load regulation than the charge pump with a high step-up ratio in [44].

A generic block diagram for a variable-frequency regulated charge pump is shown in Fig. 3.4(a). A voltage-divider ($R_{1,2}$) reduces the output voltage to within the chip's rated voltage range. The difference between this reduced voltage and the desired voltage, V_{targ}, is used to modulate the pumping frequency until the output voltage locks onto the desired value. In addition to setting the output voltage, the regulation also reduces the output resistance, increases the power-supply

Table 3.2: Charge Pump Performance

	Open loop	Closed loop
Output resistance	$R_{OL} = N/(C_p f)$	$R_{CL} = R_{OL}/K$
Power supply rejection	$\text{PSRR}_{OL} = 1/(N+1)$	$\text{PSRR}_{CL} = K\text{PSRR}_{OL}$
Start-up time constant	$\tau_{OL} = R_{OL}C_{tot}$	$\tau_{CL} = \tau_{OL}/K$

rejection, and decreases the start-up time compared to an open-loop charge pump.

The performance of an open-loop charge pump can be obtained by modeling it as the RC circuit shown in Fig. 3.4(b). From (3.1), the open-circuit voltage is $(N+1)V_{dd}$ and the open-loop output resistance is $R_{OL} = N/(C_p f)$. The total capacitance at the output, C_{tot}, combines the true load capacitance, C_L, with the distributed charge pump capacitance, $C_{eq} = NC_p/3$[34]. The open-loop performance parameters are summarized in Table 3.2.

Table 3.3: Charge Pump Variables

N	Number of stages
V_{targ}	Target voltage
I_L	Load current
R_{OL}	Output resistance of open-loop charge pump
R_F	Resistance of voltage divider
$C_{tot} = C_L + C_{eq}$	Total output capacitance
C_L	Load capacitance
$C_{eq} = NC_p/3$	Distributed charge pump capacitance
$K = K_F K_{CP}/r$	Loop gain
K_F	Voltage-to-frequency gain of the error amplifier and oscillator
$K_{CP} = I_L N/(f^2 C_p)$	Frequency-to-voltage gain of the charge pump
r	Value of voltage division

It is more difficult to determine the closed-loop regulation performance because of R_{OL}'s dependence on the operating point (i.e. R_{OL} depends upon f, which is a function of V_{out} and I_L), which makes the regulation loop nonlinear. To simplify the analysis, the small-signal model from [44] is adapted and shown in Fig. 3.4(c). The variables are all defined in Table 3.3. Sensitivity to V_{dd} has also been included in our model using the lower controlled voltage source of value $(N+1)V_{dd}$. The upper controlled source models the effect of the frequency-modulating feedback. The loop gain, K,

is the product of 1) the attenuation due to the voltage divider $(1/r)$, 2) the combined voltage-to-frequency conversion gain, K_F, of the error amplifier and oscillator, and 3) the frequency-to-voltage conversion gain of the charge pump

$$K_{CP} = \frac{dV_{out}}{df} = \frac{I_L N}{f^2 C_p} \tag{3.2}$$

To solve for the regulation performance, first equate the currents at V_{out}

$$\begin{aligned} \frac{1}{R_{OL}} \left[(N+1) V_{dd} + K \left(r V_{targ} - V_{out} \right) - V_{out} \right] = \\ = V_{out} \left(s C_{tot} + \frac{1}{R_F} \right) + I_L \end{aligned} \tag{3.3}$$

Then, solve for V_{out}, noting that by design $\frac{K}{R_{OL}} \gg \frac{1}{R_{OL}} + \frac{1}{R_F}$

$$\begin{aligned} V_{out} &= \frac{r V_{targ} + \frac{N+1}{K} V_{dd} - \frac{R_{OL}}{K} I_L}{s \frac{C_{tot} R_{OL}}{K} + 1} \\ &= \frac{r V_{targ} + \frac{1}{\mathrm{PSRR}_{CL}} V_{dd} - R_{CL} I_L}{s \tau_{CL} + 1} \end{aligned} \tag{3.4}$$

The output consists of a superposition of three components: the scaled-up target voltage, which is the desired output, as well as unwanted contributions from the supply voltage and the load current, which are both suppressed by the loop gain. The closed-loop performance parameters, which are summarized in Table 3.2, are all improved by a factor of the loop gain compared to the open-loop performance. It should be noted that the regulation circuitry adds little area and power compared to the rest of the charge pump. In section 3.1.7, we will connect these performance parameters to actual circuit parameters.

3.1.6 Current-Controlled Oscillator and Edgifier

To modulate the frequency in our variable-frequency regulator, we have used a current-controlled oscillator. In comparison to voltage-controlled oscillators, current-controlled oscillators naturally offer linear input-to-frequency gain over a wide operating range and are also easily limited so that the maximum frequency of the charge pump is not exceeded during transients. Our current-controlled oscillator, shown in Fig. 3.5, is based upon a three-stage current-starved ring oscillator. The frequency increases linearly with I_{in}. The current-to-frequency gain has been measured to be approximately 2kHz/nA over a range of 100Hz to 10MHz.

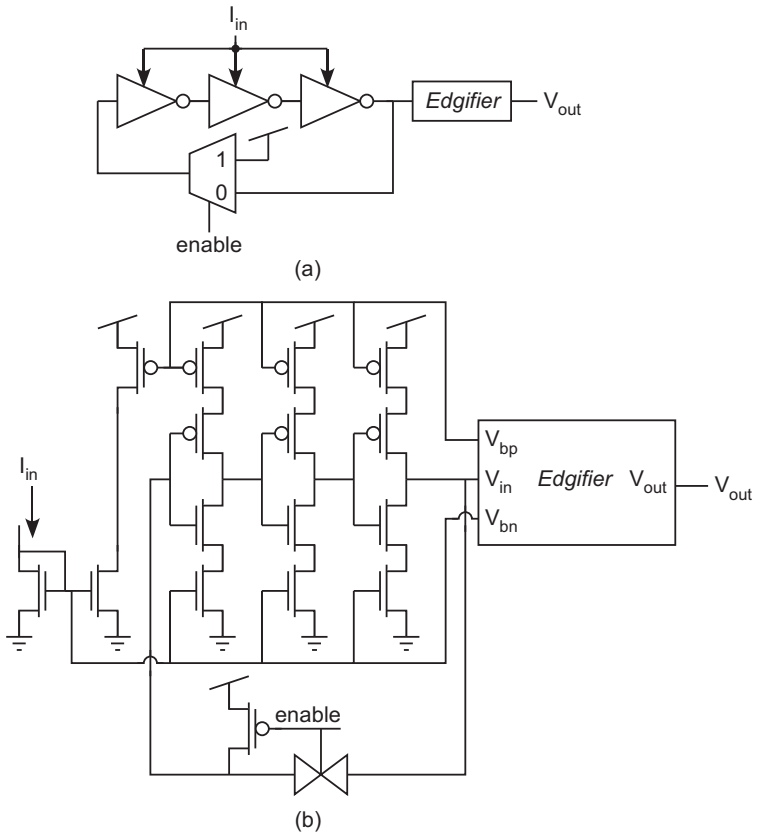


Figure 3.5: Our three-stage current-controlled oscillator with low-power edge-sharpening. (a) Block diagram. (b) Schematic.

A major attraction of variable-frequency regulation is that under light load conditions, the clock frequency reduces so that power consumption is minimized. However, the clock signals that are generated by low-frequency oscillators have a long rise and fall times. These slow-moving edges create excessive short-circuit current when they are connected directly to a subsequent logic gate. Unlike the dynamic current that charges fan-out gates, this short-circuit current performs no useful function and should not be allowed to dominate the power consumption.

To minimize short-circuit power dissipation, we control the "push" and "pull" branches of an inverter with separate non-overlapping signals. This technique is most commonly used when driving large loads—such as in clock buffers[45] or in buck converters[46]—for which it is difficult to equalize the input and output rise/fall times, and for which the consequences of short-circuit current are dire because large transistors with large current-sourcing capabilities are used. Our contribution is to adapt this concept for use with slowly-varying input signals. We call this the "edgifier" concept because it transforms slow rise and fall times into sharp edges.

Our edgifier circuit, which is shown in Fig. 3.6(a), is based upon the CMOS buffer with non-overlapping gate drive presented by Yoo[45]. Yoo's circuit consists of a push/pull buffer ($M_{8T,B}$),

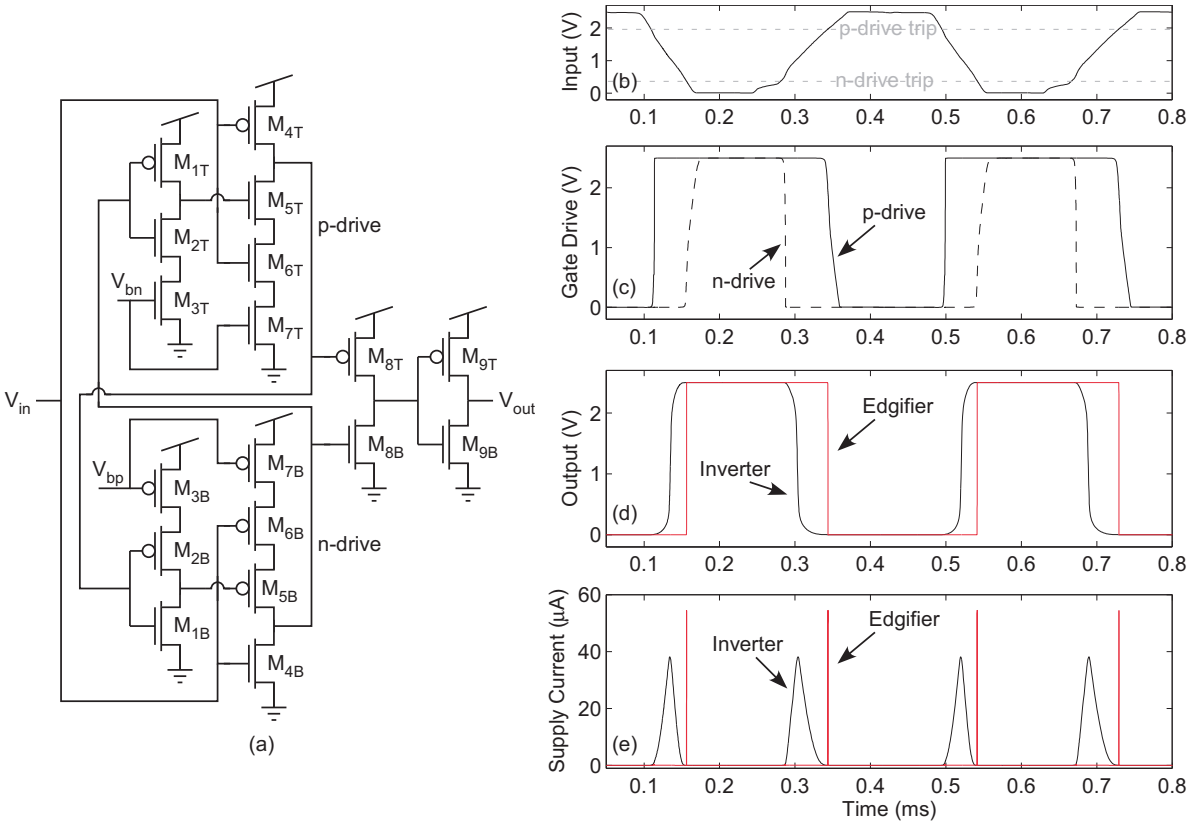


Figure 3.6: (a) Our "edgifier" circuit drives inverter $M_{8B,T}$ with non-overlapping gate signals to minimize the short-circuit current that would otherwise result from slowly rising/falling input signals. (b–e) Simulation results comparing the edgifier to a single inverter. (b) Input generated by the ring oscillator. (c) Non-overlapping gate drive signals generated by the edgifier. (d) Output signals of an edgifier and an inverter in response to the slowly rising/falling input in (b). (e) Supply current of an edgifier and an inverter. Non-overlapping gate drive significantly reduces the energy of each edgifier transition.

the gates of which are driven by the logical AND of a) the input and b) the delayed and inverted version of the complementary gate signal. As a result, one transistor is always "turned off" before the other is "turned on." This technique reduces the short-circuit current of a buffer for which the load is not pre-determined[45]. However, the non-overlapping gate drive circuitry in [45] is still subject to short-circuit current from the slow rise and fall times of an input.

To minimize the short-circuit current that is caused by the slow rise and fall times of an input, we have added current-starving transistors $M_{3T,B}$ and $M_{7T,B}$ to the circuit's gate-drive front-end. The current-starving transistors limit the short-circuit current in the front-end while allowing inverter $M_{8T,B}$ to be driven with non-overlapping signals. To enable transistors $M_{8T,B}$ to be strongly turned off, the current-starving transistors have only been used on one side of the logic ladder.

An additional benefit of asymmetric current starving is that it varies the trip point of the respective inverter. For example, the "top" inverters ($M_{1-7,T}$) have their trip points shifted toward

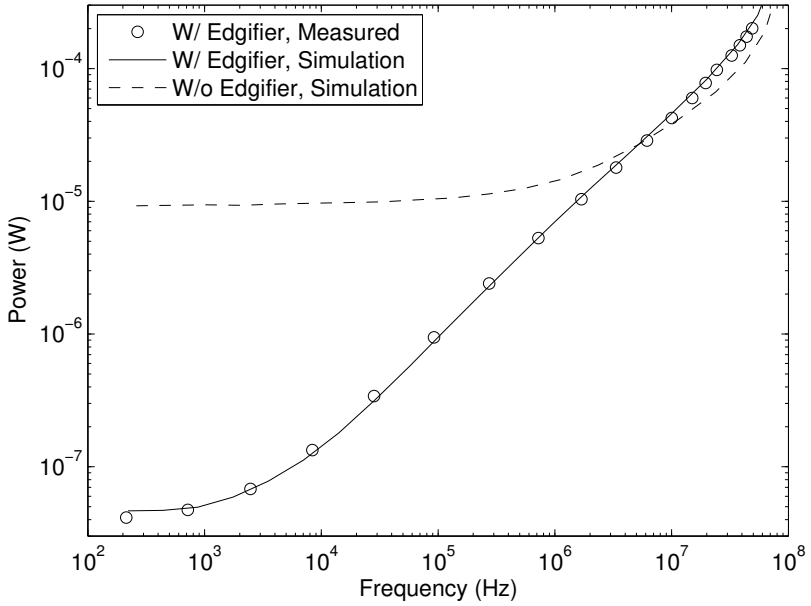


Figure 3.7: Power versus frequency of our current-controlled oscillator. The placement of an edgifier prior to any digital logic allows the power to reduce with frequency over a much larger range. The supply voltage is 2.5V, and the current-starving bias for the edgifier scales with the current-starving bias of the clock generator circuit.

higher voltages, and the "bottom" inverters ($M_{1-7,B}$) have their trip points shifted to lower voltages. The exact location of the trip point of the inverter can be tuned via the current-starving bias. For example, biasing $M_{3T,B}$ and $M_{7T,B}$ in the subthreshold domain significantly separates the "top" and "bottom" trip points while also limiting the switching current in those inverters.

As a result, our edgifier leverages the different trip points for the "top" and "bottom" paths caused by current starving to ensure that only one of $M_{8T,B}$ is turned on at any given time, as opposed to utilizing the delay properties of inverters as in [45]. The trip points of the "top" and "bottom" paths have been indicated in Fig. 3.6(b). The resulting operation of this circuit is analogous to a Schmitt trigger insofar as the output of the edgifier transitions from low-to-high at a different input voltage than it trips from high-to-low. However, the edgifier consumes significantly less power than a Schmitt trigger in this scenario of working with slow-transitioning clock signals.

Figure 3.6(b–e) compares simulation results between our edgifier and a CMOS inverter (which also consumes less power than a Schmitt trigger for this scenario) for an input signal with slow rise/fall times. In this example, the current-starving bias in the edgifier is a subthreshold current of 1nA and the outputs are unloaded. For the system-level implementation, this current-starving bias is derived as an "error signal" from a transconductor that measures the difference between the charge pump's desired and actual output voltage. The simulated output of the ring oscillator is

used as a realistic input to the circuits [Fig. 3.6(b)]. The output of the edgifier is shown in Fig. 3.6(d), and is compared to the output of an inverter in response to the same input. This inverter has the same dimensions as $M_{9T,B}$. Figure 3.6(e) shows that the inverter draws supply current over a long duration on each transition. In comparison, the edgifier's supply current is only a short impulse.

The edgifier's reduction in the short-circuit current of subsequent logic gates can significantly reduce the overall power consumption of a circuit that contains a low-frequency oscillator. This power reduction is shown in Fig. 3.7. Measured and simulated power consumption values are shown over a wide range of frequencies. The power consumption "w/ edgifier" includes the oscillator, edgifier, and one subsequent logic gate. The power consumption "w/o edgifier" includes the oscillator and one subsequent logic gate. Because of the current-starved delay elements in the oscillator, the rise and fall times of the oscillator are a constant percentage of the clock period, which results in constant short-circuit power consumption for the logic gate "w/o edgifier" below 1MHz. In contrast, the power consumption of the oscillator "w/ edgifier" continues to reduce by almost three decades, enabling the total power of the charge pump to reduce further at low load currents.

3.1.7 The Complete Charge Pump

Figure 3.8(a) shows the schematic of our complete regulated charge pump. This charge pump was fabricated in a standard n-well 0.35μm CMOS process, and the die photograph of the 230μm $\times$ 300μm charge pump is shown in Fig. 3.9. Instead of using linear resistors in the voltage divider, we used eight diode-connected pFETs, each in their own well, to reduce the overall size [Fig. 3.8(b)]. Figure 3.8(d) shows measurements of the current draw of the divider branch at different voltages. The divider branch was designed to draw 100nA–1μA over the typical operating range (V_{out} = 10V–12.5V), which is a sufficient minimum load current to prevent the clock frequency from dropping below 10kHz. This minimum clock frequency maintains stable regulation when the target load is primarily capacitive (e.g. an array of tunneling junctions) without unnecessarily wasting power.

By dividing V_{out} by a factor of eight in Fig. 3.8(a), V_{out} is thus regulated to $8V_{targ}$. The measured transfer curve from V_{targ} to V_{out} is shown in Fig. 3.8(c). From 7.5V to 16V, the average steady-state output voltage is regulated to within 1% of $8V_{targ}$ (with a small ripple voltage, which will be discussed shortly). Deviation at the high voltages is caused by the open-loop charge pump's maximum voltage $(N+1)V_{dd} = 17.5V$. In Fig. 3.8(a), error amplification is achieved by using an

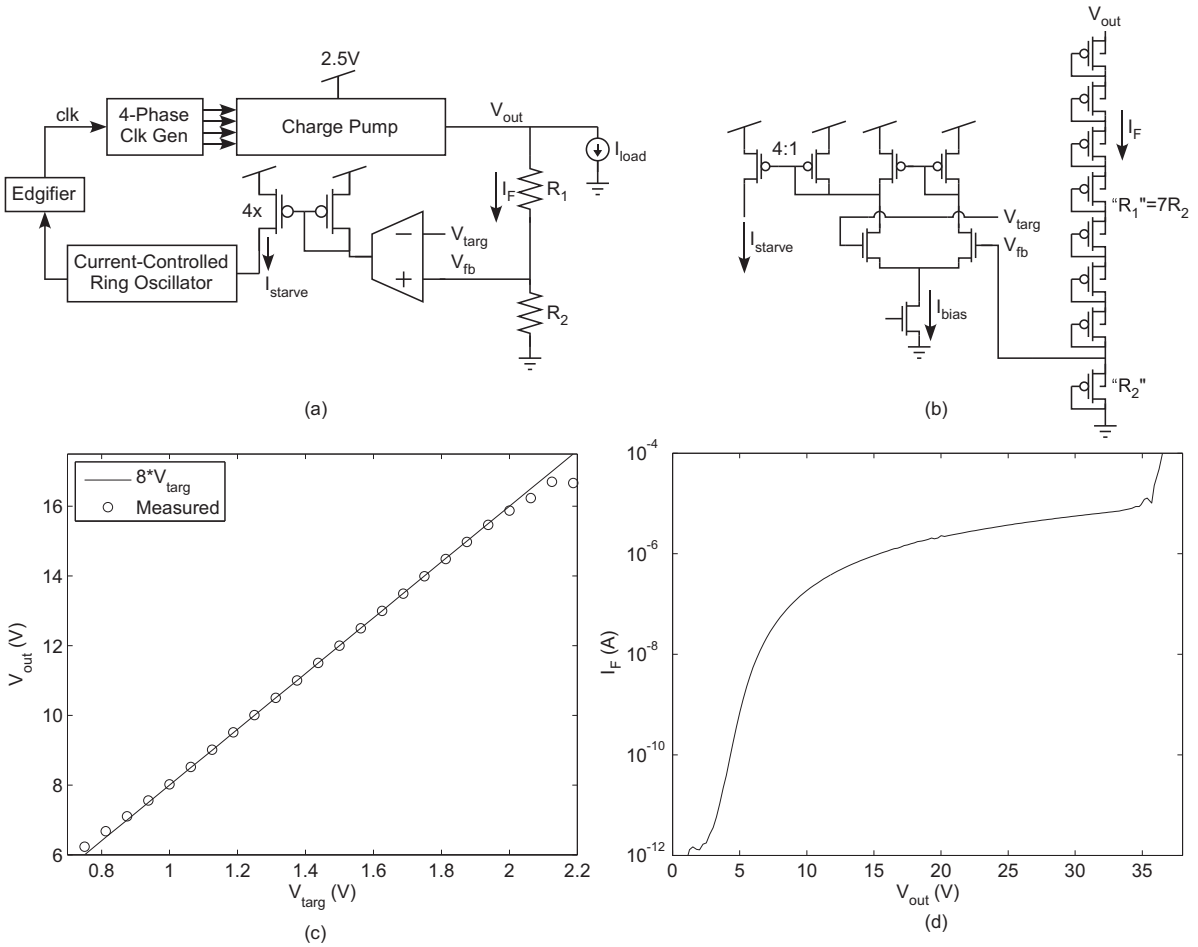


Figure 3.8: (a) Block diagram of our complete regulated charge pump. (b) Transistor-level details of the OTA and the diode-connected transistors used to implement the "resistive" divider in (a). (c) Measured dc-dependence of the charge-pump output on V_{targ} for a purely capacitive load. (d) Measured current–voltage sweep of the pFET-divider circuit. The well-to-substrate breakdown current can be seen in the top-right. This breakdown is not a concern because it occurs at a much higher voltage than the circuit's operating voltage. This result also indicates that well-to-substrate breakdown voltage is not an issue for the operating range of voltages for this charge pump.

operational transconductance amplifier (OTA) to convert the error into a current, as opposed to using the voltage output of an operational amplifier or comparator which is typically used in "skip" regulators. Deviation at the low voltages in Fig. 3.8(c) is caused by the error-amplification OTA's bias transistor being pushed out of the saturation region.

Now that the complete details of the regulated charge pump have been elaborated, we can calculate the loop gain K that was described in section 3.1.5. Starting from V_{out}: V_{out} is divided by $r = 8$, then it is converted to a current with transconductance G_m, a current mirror scales this current by a factor of 4, the current-controlled oscillator then converts this current to a frequency with a gain of $K_{RO} = 2\text{kHz/nA}$, and finally, the charge pump converts this frequency to the output

Figure 3.9: Die photograph of the complete regulated charge pump circuit. The size is 230μm $\times$ 300μm.

voltage with a gain of K_{CP}. The total loop gain is the product of all of these components

$$K = \frac{4G_m K_{RO} K_{CP}}{r} \tag{3.5}$$

The transconductance provides a way to tune the charge pump for the desired loop gain, which changes the load-regulation characteristics and the start-up time. Using (3.1), the dependence of the clock frequency on the load current is given by

$$f = \frac{NI_L}{C_p \left[(N+1)\, V_{dd} - V_{out}\right]} \tag{3.6}$$

where C_p is the pumping capacitance per stage. By combining (3.2), (3.5), and (3.6), the loop gain can be expressed as

$$K = \frac{4C_p G_m K_{RO}}{rNI_L} \left((N+1)\, V_{dd} - V_{out}\right)^2 \tag{3.7}$$

In addition to dependence upon G_m, the loop gain is also dependent upon I_L and the desired V_{out}. Of note, the loop gain increases with decreasing I_L; however, the loop gain remains finite due to the effective load current through the resistive divider in the feedback loop, thereby keeping the system stable for capacitive loads. While loop gain decreases with increasing V_{out}, it maintains a value > 100 even for an output of 16V with a purely capacitive load. Performance with capacitive loads and small I_L is an important consideration when tunneling floating-gate transistors; this charge pump is able to set $V_{out} = 8V_{targ}$ to within 1%, as was shown in Fig. 3.8(c).

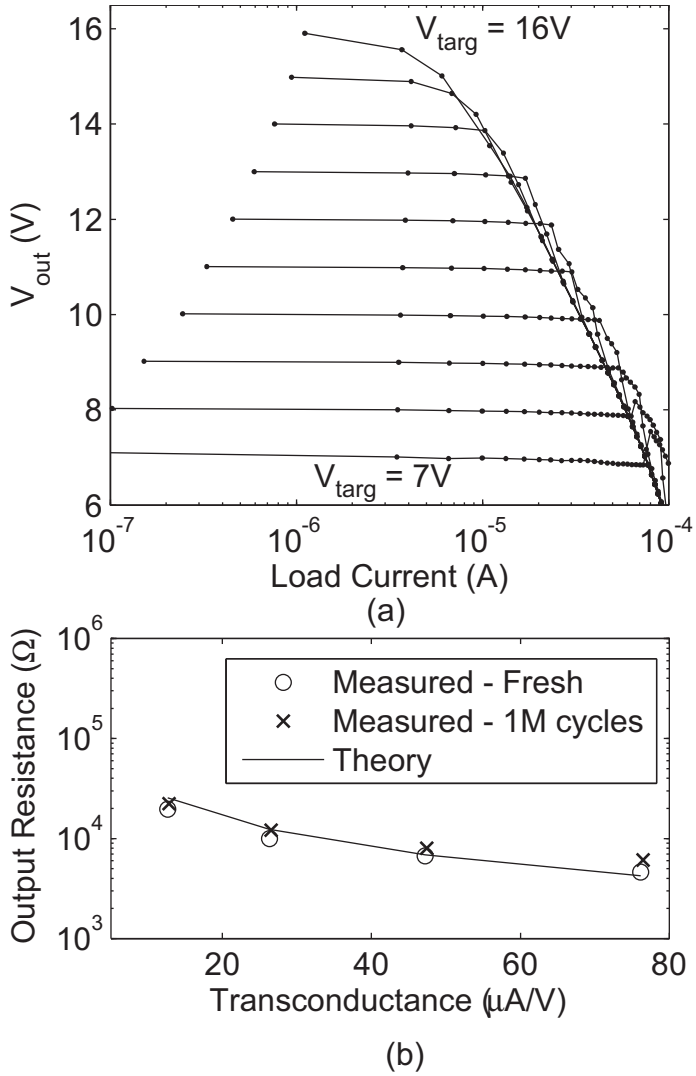


Figure 3.10: (a) Measured load regulation characteristics of the closed-loop charge pump. V_{targ} was varied from 7V to 16V in increments of 1V. (b) Measured DC output impedance of our charge pump as a function of G_m (which was varied by modifying the bias current of the OTA). To validate reliability, the measurement was performed with a fresh charge pump, as well with a charge pump that had previously generated 10^6 12.5V-pulses.

Using K and (3.1), we can obtain the closed-loop output resistance from Table 3.2

$$R_{CL} = \frac{R_{OL}}{K} = \frac{r}{4C_p G_m K_{RO}} \frac{N}{(N+1)V_{dd} - V_{out}} \tag{3.8}$$

To verify this expression, we have measured the open-loop and closed-loop load regulation in Figs. 3.3(a) & 3.10(a). The improvement afforded by regulation is clearly evident. Indeed, it would be very difficult to precisely generate an arbitrary high voltage without regulation. The output impedance is extracted from these data and is shown in Figs. 3.3(b) & 3.10(b). Good agreement is found between the measured results and the theoretical values for R_{OL} and R_{CL}. This agreement confirms that, when the charge pump is designed to sufficiently approach ideal characteristics, this

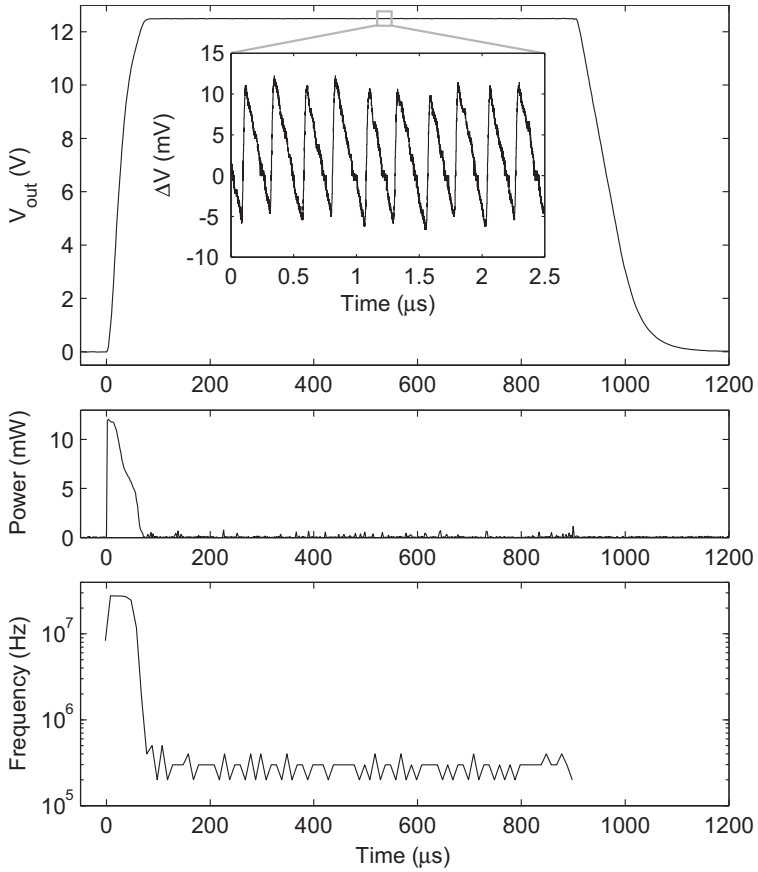


Figure 3.11: Measured transient characteristics of our closed-loop charge pump. (Top) Output voltage with an inset of the ripple voltage, ΔV. (Middle) Instantaneous power consumption. (Bottom) Closed-loop adapted clock frequency.

simple analysis can be used to confidently design a high-voltage charge pump.

Additionally, (3.8) helps provide insights into the stability of this system. Since R_{CL} has no dependence on the load current, the corner frequency of this system only depends on G_m (which is generally set to a fixed value) and the desired V_{out}. Decreasing V_{out} increases the bandwidth of the system by modifying the output pole, which is the dominant pole of the charge pump. For V_{out} within typical values for generating tunneling voltages (e.g. > 7.5V), all non-dominant poles are at sufficiently high frequencies so as to not cause any stability issues, whatsoever. For values of $V_{out} < 7.5$V where instability could become an issue, stability can be maintained with this topology by increasing C_L to keep the dominant pole sufficiently low, or by reducing the number of stages.

In a circuit that operates beyond the rated voltage of the process, the designer should ensure that the local voltage differentials for each device are within the rated voltage range. A beneficial feature of the charge pump stage [Fig. 3.2(a)] is that the use of two series switches in each stage protects the devices from any voltage stress greater than V_{dd}[36]. To verify that this protection

ensures reliable performance under typical operating conditions, we measured the charge pump's output resistance before and after the charge pump had generated 10^6 12.5V-pulses of 1ms duration. These pulses are typical of the way the charge pump is used to program floating-gate transistors. The before-and-after measured output resistance is shown in Fig. 3.10(b). The "burned in" charge pump consistently has a slightly higher output resistance. However, the variation is small, and the number of cycles is greater than the typical rating for Flash memory, which confirms that this charge pump has sufficient long-term reliability for our application.

The prominent characteristic of a frequency-regulated charge pump is that the frequency varies, which helps to minimize the power consumption once the target output voltage is reached. Figure 3.11 shows a measurement of the charge pump generating a 1ms, 12.5V tunneling pulse. The measured frequency of the clock over time is shown in Fig. 3.11(c). During startup, the OTA is saturated and the clock pumps at a maximum frequency of ~30MHz. Once the target voltage is reached, the clock is relaxed to ~300kHz. The resulting mitigation in supply current while the voltage is held is seen in Fig. 3.11(b) which is especially pronounced because the clock frequency is slow enough that the edgifier produces additional power savings (see Fig. 3.7). The overall energy that was used to generate this pulse was 1.45μJ.

Figure 3.11(a) also shows the output voltage ripple, ΔV, for the charge pump in steady-state. The standard expression for output ripple in charge pumps is given by

$$\Delta V = \frac{I_L \Delta t}{C_L} \tag{3.9}$$

where C_L is the capacitance loading the output node, I_L is the load current that discharges C_L, and Δt is the period of the ripple which is established by the clock frequency; for our charge pump that uses a parallel structure, Δt is half the clock period (i.e. twice the clock frequency). Since our charge pump leverages variable-frequency regulation, the frequency of the clock depends on I_L as given by (3.6). Therefore, the total output ripple voltage for our charge pump is

$$\Delta V = \frac{C_p}{2NC_L} \left[(N+1) V_{dd} - V_{out} \right] \tag{3.10}$$

Of note, the ripple voltage of our charge pump has no dependence on the load current or the clock frequency, whereas a charge pump that does not employ variable-frequency regulation has these dependencies (see (3.9)). Instead, the ripple voltage can be designed to have a specific value

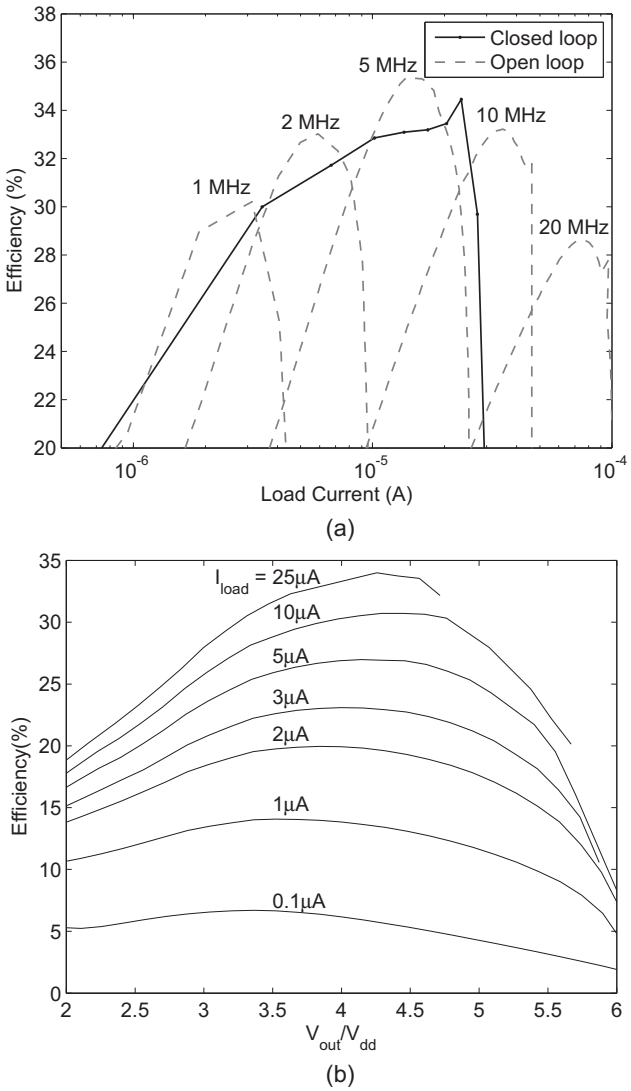


Figure 3.12: Measured efficiency of the charge pump. (a) Measured open-loop and closed-loop efficiency for varying load currents. The closed-loop charge pump was measured with $V_{out} = 12$V. (b) Measured efficiency for varying step-up factors and multiple load-current conditions.

by setting appropriate values for the pumping capacitance and the load capacitance. Our charge pump had a measured ripple of $< 18mV$ when driving a load capacitance of approximately $80pF$, which was created primarily from parasitic capacitances from the pads, board-level connections, and oscilloscope probes (i.e. no explicit load capacitor was used).

The efficiency of a charge pump is the power delivered at the output of the charge pump divided by the total power going into the circuit, given by

$$\gamma = \frac{V_{out} I_L}{V_{dd} I_{vdd}} \tag{3.11}$$

For our regulated charge pump, this input power includes the power consumed by all components,

Table 3.4: Comparison of Regulated High-Voltage Charge Pumps

	This Work	Aaltonen [44]	Barnett [47]	Kang [40]	Kim [48]	Tsai [49]	Tsai [50]	Tseng [51]
Process	0.35μm	0.35μm HV	0.35μm	63nm	0.13μm	0.18μm	0.18μm	0.18μm
V_{in} (V)	2.5	2.5	2–3	1.8	1.2	1.8	1	1.2
V_{out} (V)	7.5–16	10	13.2–14.1	1–18	3	±6	3–6	6
Max Step-Up	6.4	4	7.05	10	2.5	3.33	6	5
N	6	9	8	45	3	3	6 stages by 9 interleaved cells	3
Effective Step-Up per Stage	1.07	0.44	0.88	0.22	0.83	1.11	1	1.67
Total C_{pump} (pF)	18	14.4	–	225	240	–	–	–
Size (mm^2)	0.069	0.14	0.092	>2.56	0.6	4.94	0.5	4
Efficiency	34% @ 25μA	18% @ 29μA	–	–	12.5% @2mA	–	48–58% 52% @ 240μA	–
Maximum Load Current	40μA @10V	–	<5μA	150μA	2mA	400μA	240μA	700μA
Ripple (mV)	18	–	–	200	–	–	39	30
Tunneling Possible?	Yes	Yes	Yes	Yes	No	No	No	No

not just the charge pump. We have not emphasized efficiency because it is not a crucial specification when generating short tunneling pulses for tunneling junctions that draw a very small load current. As can be seen in Fig. 3.11, most of the energy is consumed while starting up the charge pump. However, we will briefly discuss efficiency because it is a standard comparison point for voltage converters and because it will be of interest for modifying this charge pump for other applications, such as for injection-level supply voltages for floating-gate programming and also for various MEMS applications.

From [34], the supply current of an ideal charge pump with bottom-plate stray capacitance is

$$I_{vdd} = \left[(N+1) + \alpha \frac{N^2}{(N+1)V_{dd} - V_{out}} V_{dd}\right] I_L \tag{3.12}$$

where α is the ratio of the bottom-plate stray capacitance to the pumping capacitance, which is fixed for a given CMOS process and is typically in the range of 0.1–0.2. The first additive term accounts for the current that is pumped toward the load. The second term accounts for the current that charges and discharges the stray capacitance. The theoretical maximum efficiency is

$$\gamma = \frac{V_{out}}{V_{dd}\left[(N+1) + \alpha \frac{N^2}{(N+1)V_{dd} - V_{out}} V_{dd}\right]} \tag{3.13}$$

For V_{out} = 12V, N = 6, and V_{dd} = 2.5V, the maximum theoretical efficiency is approximately 52%. Figure 3.12(a) shows the measured efficiency of the open-loop and closed-loop charge pump

for varying load currents. Figure 3.12(b) also shows how the efficiency changes according to the step-up factor (for multiple values of I_L). The charge pump achieves approximately 65% of the theoretical efficiency. Furthermore, the overhead of regulation has not significantly decreased the efficiency of the unregulated charge pump. In fact, variable-frequency regulation is able to achieve better regulation across a wider range of load currents.

Since the charge pump was designed for use in ultra-low-power systems where the supply voltage is provided by batteries or energy harvesting and may vary significantly over time, power-supply rejection is an important concern. The use of regulation improves the power-supply rejection by 68dB over an open-loop charge pump. The measured power-supply rejection ratio (PSRR) for the regulated charge pump was 52dB at 1kHz with $V_{out} = 12$V.

Table 3.4 compares our charge pump to others—specifically, we compare our charge pump to other regulated charge pumps that have been fabricated and are capable of producing voltages that are twice the rated voltage of the core devices (i.e. $V_{out} > 2V_{dd}$). The last row of this Table indicates whether or not the particular charge pump can achieve large-enough voltages for tunneling floating-gate transistors in their respective process. Even though a constant and predictable charge-pump output voltage is critical for many memory applications (e.g. multi-level Flash [52] and analog non-volatile memory [31]), very few regulated high-voltage charge pumps have been reported. As can be seen from this Table, our charge pump is able to provide good performance while maintaining a small size. Additionally, our charge pump is able to provide a high voltage with very little ripple, and (3.10) describes how this ripple can be improved even further.

3.2 A Regulated Charge Pump for Injecting Floating-Gate Transistors

In the previous section, we presented a charge pump capable of generating the voltages required for programming, and we focused on tunneling which is the larger voltage that must be generated. We were able to show that the charge pump was able to produce high voltages with relatively low ripple in the output voltage. Tunneling is reserved for global erasure in many analog FG applications [53], and so a moderate amount of ripple on the tunneling voltage is acceptable. However, injection is often used for precise programming, and thus the output ripple should be very small (on the order of 1mV or less).

In this section, we extend our previous results in [3] to achieve extremely low ripple in the

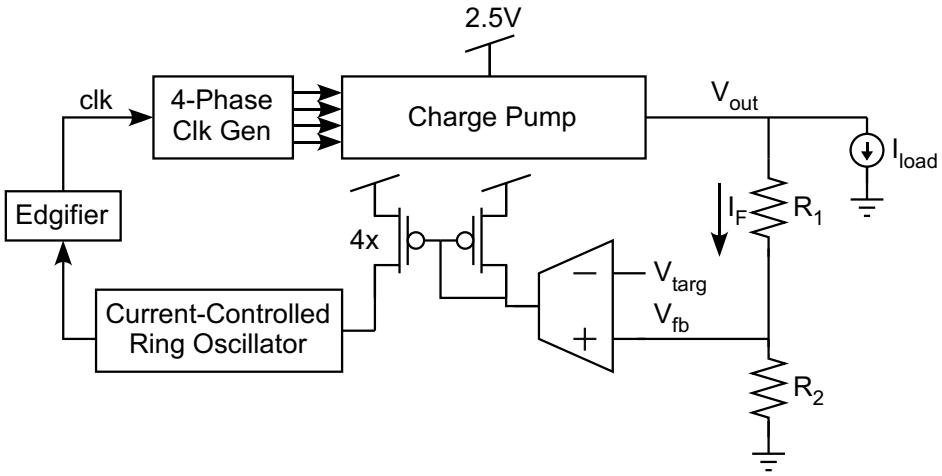


Figure 3.13: Block diagram of tunneling charge pump we presented in [3].

output voltage to provide voltages for injection. We present a method to reduce output ripple in high-voltage charge pumps by adding a high power-supply-rejection-ratio (PSRR) low drop-out regulator (LDO) in the regulation loop of the charge pump. This $260\mu m \times 460\mu m$ charge pump was fabricated in a standard 0.35μm CMOS process. While operating from a 2.5V supply, the charge pump generates regulated voltages from 3V to 10V. The maximum efficiency of the charge pump is 23% for 25μA of load current. The output ripple is less than 1mV for a wide range of load currents and output voltages. We describe the development of this regulated charge pump in the remainder of this Chapter.

3.2.1 Considerations Regarding Injection Voltages and Scaling

Channel hot-electron injection is used to precisely set a specific charge on an FG in many programmable analog systems [17]-[18]. In Chapter 2 considerations regarding injection voltages and scaling were discussed. We showed how the voltages needed for injection ($V_{sd,inj}$) scale with technology nodes. Charge pump circuits are used to multiply the chip supply voltage by a constant and generate the voltage required for programming. A charge pump used for programming is enabled for a short period of time to minimize the energy consumption. The input to the charge pump is the chip supply voltage. Since the programming infrastructure [2] requires an additional overhead of several hundred millivolts beyond the necessary V_{sd} for injection, a charge pump for injection must be able to provide $(2 \sim 3) \times V_{dd}$. In addition, in order to have a suitable charge pump for the injection process, we must have a circuit that has low output ripple. Some other aspects of a good charge pump are fast start-up time and small die area.

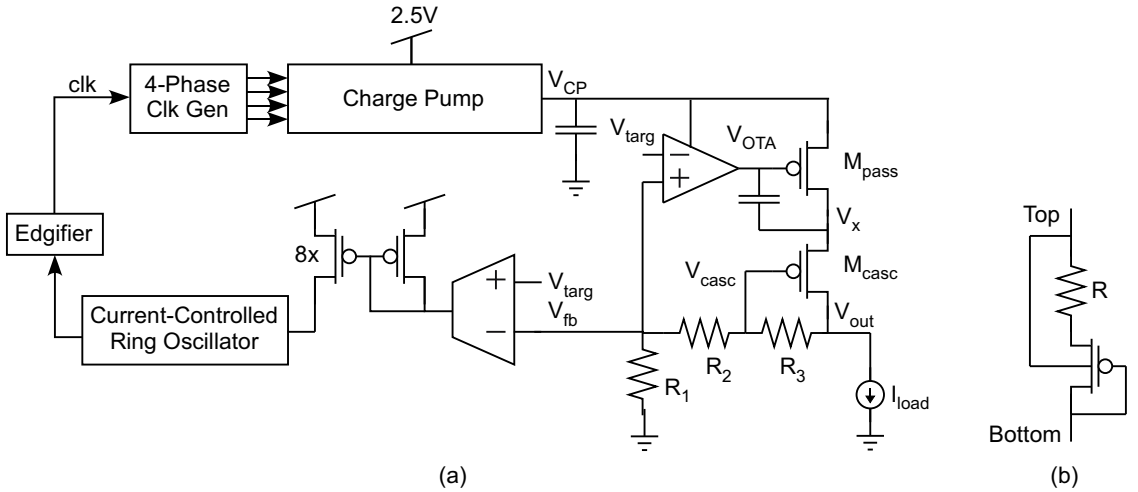


Figure 3.14: (a) Block diagram of the proposed injection charge pump. (b) Transistor-level details of the resistor cell to implement the resistive divider.

Table 3.5: Charge pump specifications.

# stages	6
C_{bt}	110fF
C_p	3pF
M_{bb}	3μm x 0.35μm
M_{bt}	3μm x 0.35μm
M_{sw}	5μm x 0.35μm

3.2.2 Proposed High Voltage Charge Pump

Figure 3.13 shows the block diagram of the charge pump we presented in [3]. A voltage divider shifts the output voltage to a voltage between ground and V_{dd}. The difference between this feedback voltage (V_{fb}) and the target voltage (V_{targ}) is converted into a current. This current modulates the frequency of the current-controlled ring oscillator. The open-loop charge pump used in this work includes 6 stages which generate a high-voltage that is related to the oscillation frequency of the oscillator. The circuit schematic of the charge transfer stage (CTS) is shown in Fig. 3.15(a) [3]. Figure 3.14(a) shows the circuit diagram of the proposed high-voltage charge pump. The charge pump presented in [3] has a small ripple (around 20mV) which is adjustable with the load capacitor. In this work, we used an LDO inside the loop of [3] to make the ripple smaller. We did not place the LDO outside the charge pump loop because that resulted in larger ripple and longer settling time.

The operation of the closed-loop system of Fig. 3.14(a) is as follows. Starting from V_{out}: V_{out} is divided by r = 5, then it is converted to a current with transconductance G_m, a current mirror scales this current by a factor of 8, the current-controlled oscillator then converts this current to a

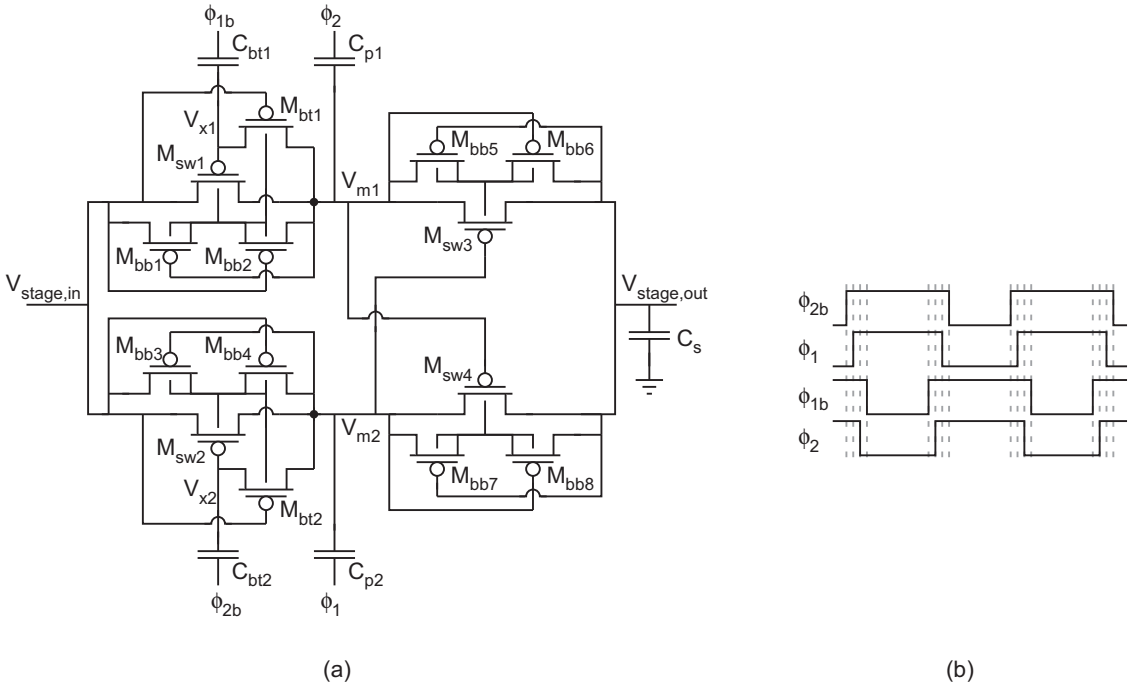


Figure 3.15: (a) Circuit diagram of the charge transfer switch (CTS). (b) Non-overlapping clock signals necessary for the CTS.

frequency with a gain of $K_{RO} = 2kHz/nA$, and finally, the charge pump converts this frequency to the output voltage of the charge pump (V_{CP}). The design specifications of the charge transfer switches used in this work are provided in Table 3.5. V_{CP} has big ripple which makes it unsuitable for precise programming; however it is used as the supply voltage of the low drop-out regulator (LDO). The sizes of the resistors in the LDO voltage divider are set to: $R_1 = R_2/3 = R_3$. Since the voltage at the negative input of the operational transconductance amplifier (OTA) inside the LDO loop is at V_{targ}, the LDO will regulate the output voltage to be $5V_{targ}$.

The pass transistor used in this charge pump is a series connection of two pMOS transistors. This structure is similar to the high-PSRR LDO presented by [54]. This cascode structure improves the output impedance of the pass transistor, which leads to an improved supply rejection. As mentioned in [54], this structure will reduce the gate capacitance associated with the pass transistor and will make compensation of the LDO easier. Thick-oxide devices are used for M_{pass} and M_{casc}, since the voltage across these transistors can be large and may damage the devices. Using the saturation region condition for pMOS transistors and knowing that the voltage at the gate of the M_{casc} is at $V_{casc} = 4V_{out}/5$, the maximum voltage allowed at the output to keep M_{casc} in saturation region is:

$$5|V_{Th,thick}| > V_{out} \tag{3.14}$$

where $V_{Th,thick}$ refers to the threshold voltage of the thick-oxide transistors. Since the body of

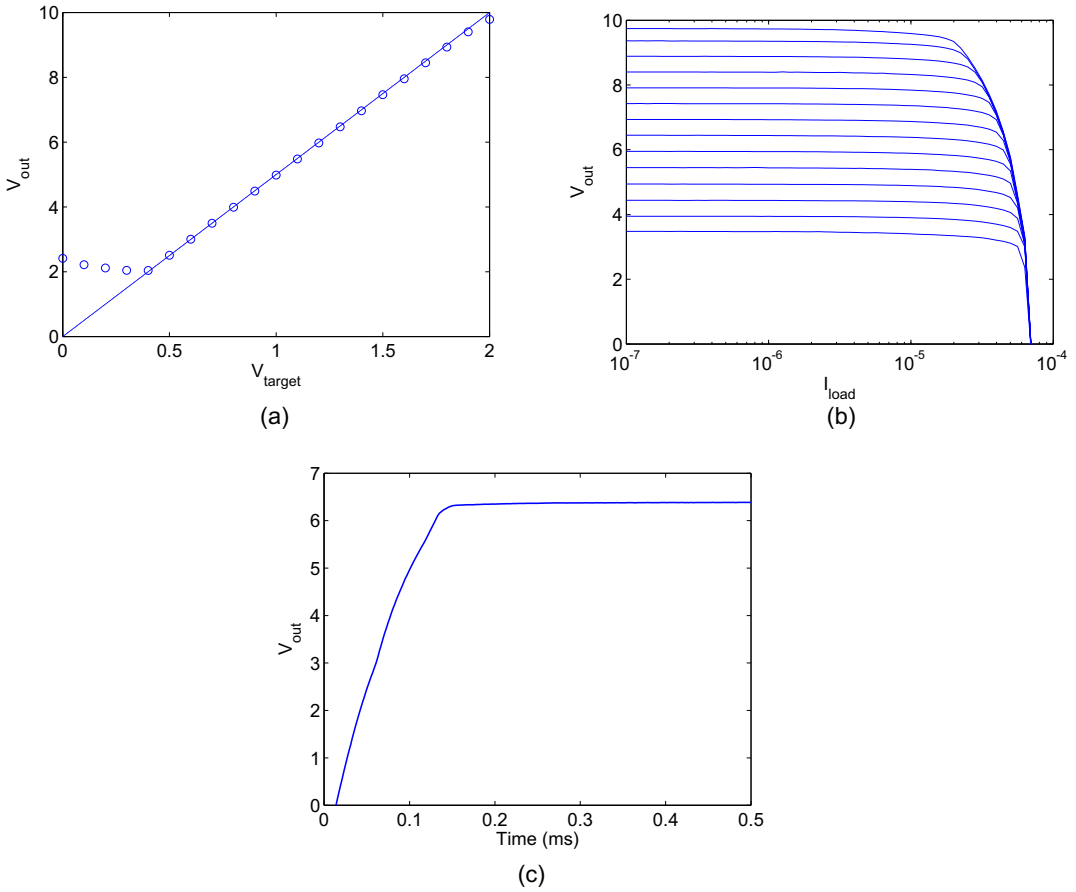


Figure 3.16: (a) Measured DC dependence of the charge-pump output on V_{targ} for a purely capacitive load. (b) Measured load regulation of the closed-loop charge pump for various output voltages (c) Measured transient response of the closed-loop charge pump for a 20μA load current.

M_{casc} is connected to V_{CP}, the voltage difference between the body and source of this transistor (V_{bs}) will increase the threshold voltage of M_{casc}. Hence, the upper limit ($5V_{Th,thick}$) on the output voltage increases, too. The range of voltages used for the injection process in 0.35μm CMOS is between 5V and 6.5V; this upper limit ($5V_{Th,thick}$) is high enough to cover the injection voltage range. The saturation region condition for the pass transistor (M_{pass}) can be expressed by:

$$V_{CP} > \frac{4V_{out}}{5} + V_{Th,casc} + \sqrt{\frac{2I_{load}}{K_{pass}}} + \sqrt{\frac{2I_{load}}{K_{casc}}} \tag{3.15}$$

where $K_i = \mu C_{ox}(W/L)_i$ is the aspect ratio of M_i, $V_{Th,casc}$ is the threshold voltage of the cascode transistor, and I_{load} is the load current.

According to (3.15), for a specific output voltage (V_{out}) and load current (I_{load}), M_{pass} can be biased in the saturation region by correct selection of K_{pass}, K_{casc}, and the open-loop charge-pump parameters to set V_{CP}. The output voltage of the open-loop charge pump can be set by the following equation [34]:

$$V_{CP} = (N+1)V_{dd} - N\frac{I_{load}}{C_P f} \tag{3.16}$$

To make V_{CP} satisfy (3.15), the number of charge transfer stages (N) and the size of the pumping capacitors (C_p) must be adjusted correctly. The operational frequency of the charge pump (f) is a function of I_{load} and V_{out} and will be adjusted by the regulating operation of the charge pump.

Figure 3.14(b) shows a single resistive cell used in the voltage divider branch of the LDO. Each cell includes a resistor ($98k\Omega$) in series with a pFET, all in a single well, to reduce the overall size of the resistor divider circuitry. R_1 and R_3 are each a single resistive element, and R_2 is a series of three of the resistive elements as shown in Fig. 3.14(b). The resistance seen between the top and bottom terminals of this resistive cell can be expressed as:

$$R_o = \frac{(g_m + g_{mb})R + \frac{R}{r_o} + 1}{g_m + \frac{1}{r_o}} \tag{3.17}$$

3.2.3 PSRR of LDO

In this section, we establish a frequency-domain equation for the PSRR of the proposed LDO based on [55]. When analyzing only the LDO of Fig. 3.14(a), a small-signal input voltage, V_i, will induce an output voltage, V_o. The ratio between V_o and V_i simply represents the LDO PSRR. The PSRR of the proposed LDO can be expressed as:

$$\frac{V_o}{V_i} = \frac{5(g_{m2}g_{ds1} + g_{m2}g_{m1} + g_{mb}g_{m1} + g_{ds2}g_{ds1} + g_{m1}g_{ds1})(1 + R_aC_as)}{(g_{m2} + g_{mb} + g_{ds2})g_{m1}A_{ota} + \frac{5(g_{m2}+g_{mb}+g_{ds2})(1+R_aC_as)}{z_L} + (4g_{m2}g_{ds1} + 5g_{ds2}g_{ds1})(1 + R_aC_as)} \tag{3.18}$$

where Z_L can be expressed as the parallel combination of load capacitance (C_L) and the total resistance of the resistor divider ($5R_o$):

$$Z_L = \frac{5R_o}{1 + 5R_oC_Ls} \tag{3.19}$$

Assuming that (g_{m1}, g_{m2}, g_{mb}) $\ll$ (g_{ds1}, g_{ds2}) and that the resistance seen by the resistor divider ($5R_o$) is high, the PSRR at DC can be approximated as:

$$\frac{V_o}{V_i}|_{DC} = \frac{1}{\frac{g_{m1}A_{ota}}{5g_{ds1}} + 1} \tag{3.20}$$

Hence, by having a high-gain OTA (A_{OTA}) and keeping M_{pass} in saturation, the denominator of (3.20) will have a large value, thus significantly improving the PSRR.

3.2.4 Experimental Results

This charge pump was fabricated in a standard n-well 0.35μm CMOS process, and the die area of the charge pump is $260\mu m \times 460\mu m$. A conventional 5 transistor differential amplifier was employed for both error amplifiers. The OTA used inside the charge-pump loop has a pFET input structure, and the OTA used inside the LDO loop has a nFET input structure. The bias currents

of the OTAs inside the charge-pump loop and the LDO loop are 270nA and 4μA, respectively. The measured transfer curve from V_{targ} to V_{out} is shown in Fig. 3.16(a). Deviation at the high voltages is caused by the maximum limit that is provided by (3.14) and (3.15). Deviation at the low voltages in Fig. 3.16(a) is caused by the error-amplification OTA's input transistors in the LDO being pushed out of the saturation region. The closed-loop load regulation of the charge pump is shown in Fig. 3.16(b). The injection process requires that the output impedance is designed such that the output of the charge pump stays constant in a wide range of load currents. Hot-electron injection requires voltages higher than V_{dd} to make electrons pass through the gate oxide. For our application, we will mostly use 6.5V as the elevated supply voltage in the injection process.

Figure 3.16(c) shows the measurement results of the charge pump generating a 6.5V injection pulse with 20μA of load current for a load capacitance of 0.47μF. This plot shows that the output voltage ramps up to the target voltage in less than 0.2ms.

The efficiency of a charge pump can be expressed by:

$$\gamma = \frac{V_{out} I_{load}}{V_{dd} I_{vdd}} \tag{3.21}$$

which is the power delivered to the load divided by the total power going into the charge-pump circuit. Figure 3.17(a) shows the efficiency measurement of this charge pump for 6.5V output. The maximum efficiency is around 23%, and it starts to drop for load currents greater than 25μA because the loop gain is not high enough to hold the output voltage at 6.5V. Figure 3.17(b) shows the efficiency measurement of this charge pump for different step-up ratios (V_{out}/V_{dd}) for 20μA of load current. This plot shows that this charge pump is able to provide higher voltages with better efficiency.

The accuracy of the injection process is a function of the accuracy of the elevated injection voltage (e.g. 6.5V in our application). The required accuracy for programming will set a maximum limit on the ripple voltage (on the order of 1mV or less). Because the ripple was very small and within the limits of our measurement equipment, we used RC extracted simulations to report the ripple amplitude. Figure 3.18 shows the RC-extracted simulation results when the charge pump was set to 6.5V output. The load current was set to 0μA and 20μA in these simulations, and the simulated peak-to-peak output ripple is approximately 0.4mV and 0.6mV, respectively.

3.3 A high voltage charge pump using a feed-forward compensation technique

In this section, we present a charge-pump for injecting FG transistors. A few modifications are made to the injection charge-pump presented in section 3.2. I will explain these modifications in the remainder of this section.

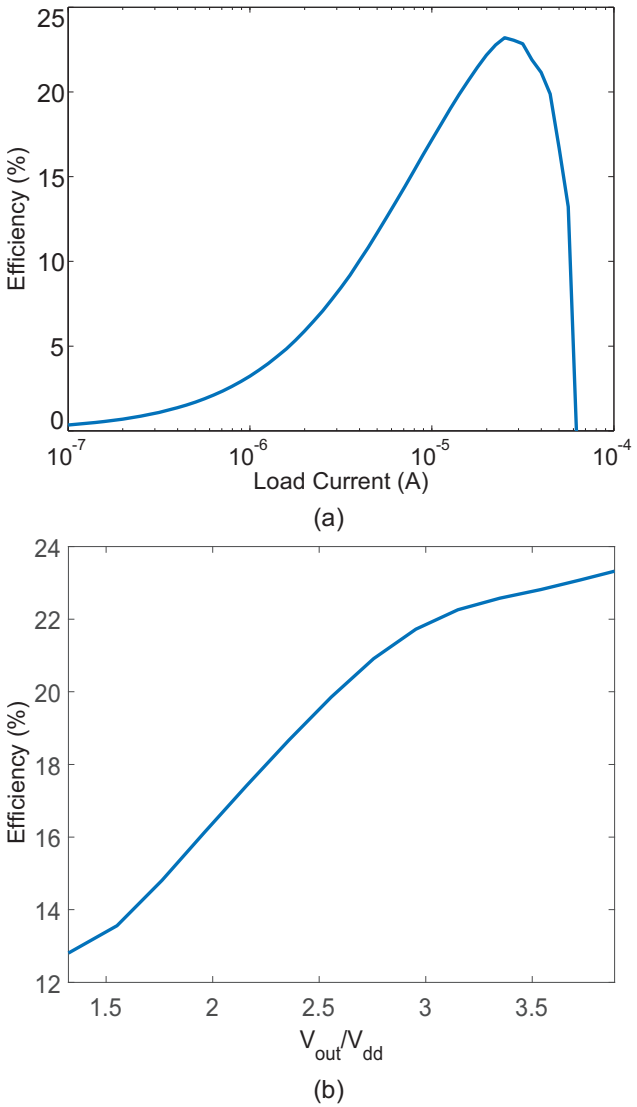


Figure 3.17: (a) Measured efficiency of the charge pump versus load current when $V_{out} = 6.5V$ (b) Measured efficiency of the charge pump versus V_{out}/V_{dd} for 20μA of load current.

First, a higher number of charge-transfer stages (e.g. 7) are used in this structure to improve the DC level at the input of the LDO. This will allow the pass transistor and the cascode transistor operate in the above threshold saturation region.

Second, since the OTA inside the LDO loop consumes a portion of the current provided by the open loop charge-pump, we must make sure that enough current will be provided by this open loop charge-pump. Thus, two different sizes for pumping capacitors are tried on the fabricated chip (3.14pF and 4.72pF). I use bigger pumping capacitors to improve the current provided by the open loop charge-pump.

Third, the body of the cascode transistor is tied to the source of the same transistor. According to (3.15), this allows the pass transistor to operate in the above threshold saturation region.

Finally, a feed-forward compensation capacitor is used in parallel with the resistor divider of the LDO to improve the phase margin of the LDO. Thus the output voltage will settle down to $5 * V_{targ}$

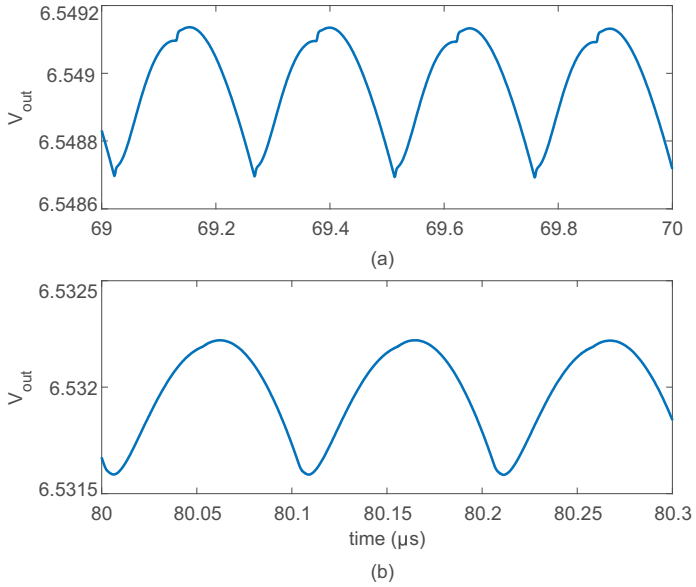


Figure 3.18: Simulated ripple for an output voltage of 6.5V with (a) 0μA load current and (b) 20μA load current.

with an over-damped behavior. This is very important for the injection of the FG transistors. The amount of injection current in a PMOS device can be represented by the following equation:

$$I_{inj} \approx \beta {I_s}^{\alpha} e^{V_{sd}/V_{inj}} \tag{3.22}$$

where α, β, and V_{inj} are fit parameters. This equation covers the subthreshold operating range, which is where the injection operation is most efficient. Of note, the channel current (I_s) and the source-to-drain voltage (V_{sd}) determine the injection rate. As was mentioned earlier, significant injection only occurs for $V_{sd} > V_{dd}$. According to (3.22), the injection current of an FG transistor is exponentially related to the drain-source voltage of the FG transistor. Thus, any overshoot at the injection voltage provided by the charge-pump during injection can cause a spike in the injection current. Therefore, that can cause inaccuracy in the programmed charge on the floating node. The details of the proposed injection charge-pump circuit will be presented in the rest of this section.

3.3.1 Low ripple charge-pumps

There are two available approaches to generate a high DC voltage using a closed loop charge-pump and an LDO. First, using a closed loop charge-pump in series with a closed-loop LDO. Second, using a closed loop LDO inside a closed loop charge-pump. The first circuit is presented in Fig. 3.19. The closed loop charge-pump uses 7 charge-transfer stages. The ratio of R_3/R_4 is set to 7 in this closed loop charge-pump. The LDO uses the same exact structure as the injection charge-pump proposed in section 3.2. Fig. 3.20 shows the simulation results of the charge-pump under a 50μA of load current. The output voltage of the LDO is shown in Fig. 3.21. The output

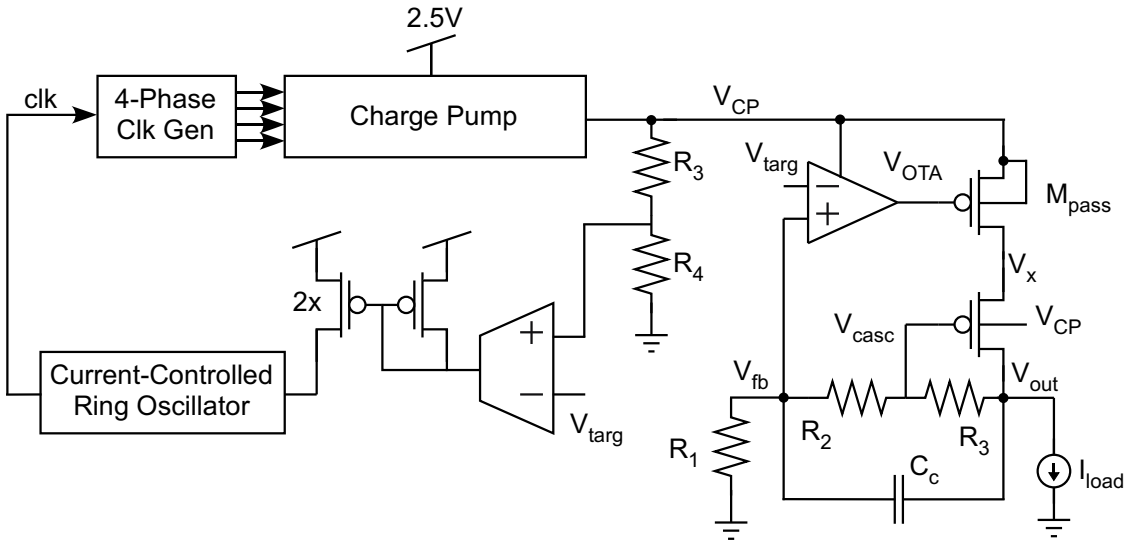


Figure 3.19: Circuit structure of the proposed open loop charge pump in series with an LDO

voltage of the charge-pump in this case overshoots to a very large value (approximately 8V) and settles down to the target voltage (6.5V). This behavior is not desirable in FG programming, since the FG transistor will be over injected during the large overshoots. I would like to get an over-damped output voltage for FG programming. In the next section, I will show that using an LDO inside a charge-pump loop will provide an over-damped behavior.

3.3.2 Using LDO inside the charge-pump loop

In order to have an accurate FG injection, It is required to generate a low noise high-voltage signal using the proposed charge pump. The high voltage DC signal generated from the open loop charge pump has a considerable ripple. According to equation (3.10), the amplitude of this signal (V_{CP}) is controllable with C_{CP}. Thus, we will need to integrate a big capacitor on the chip (C_{CP}) or use an external pin to tie an off-chip capacitor to the output of this charge-pump to generate a low ripple output voltage. By using an LDO inside the charge-pump loop we can reduce the size of this capacitor (C_{CP}). Thus, all elements of this charge-pump can be integrated on the same chip.

The circuit diagram of the proposed LDO is shown in Fig. 3.22-(a). The condition to keep the pass transistor inside the saturation region is:

$$V_{CP} > \frac{4V_{out}}{5} + V_{Th,casc} + \sqrt{\frac{2I_{load}}{K_{pass}}} + \sqrt{\frac{2I_{load}}{K_{casc}}} \tag{3.23}$$

Using a body to source connected cascode transistor (M_{casc}), the threshold voltage of the cascode transistor will be at the lowest possible value. Thus compared to the charge-pump presented in section 3.2, the load current (I_{load}) on the right side of (3.23) can be at a higher level and still keep the pass transistor inside the above threshold saturation region. This means that the charge-pump

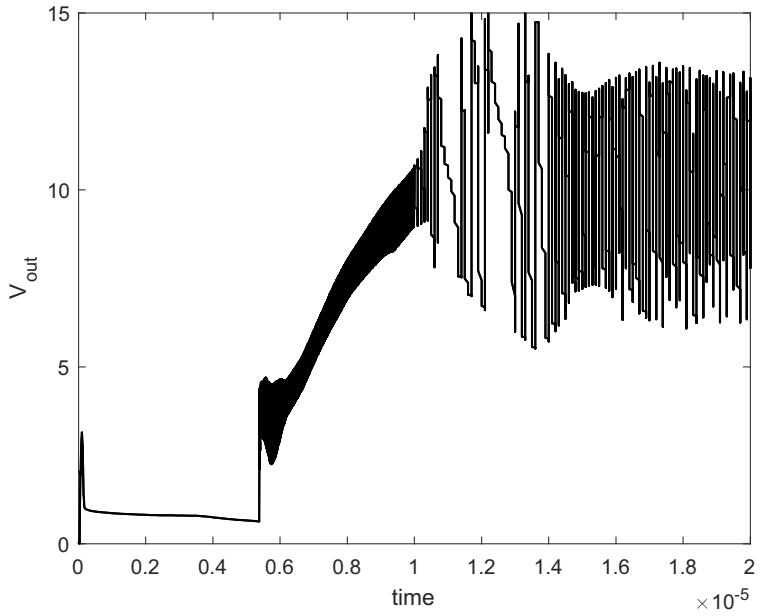


Figure 3.20: Output of the charge-pump (V_{CP}) presented in Fig. 3.19

can handle higher load currents compared to the injection charge pump presented in section 3.2.

Another important structural change to the proposed circuit is using a feed-forward capacitor inside the LDO loop to improve the loop stability of the LDO [56]. A capacitor can be added to the original LDO configuration, to provide a high-frequency bypass path for the loop gain. This capacitor generates a pole-zero pair in the open-loop transfer function as follows [56]:

$$H(s) = \frac{A_0(1 + \frac{s}{w_{z1}})}{(1 + \frac{s}{w_{p1}})(1 + \frac{s}{w_{p2}})(1 + \frac{s}{w_{p3}})} \tag{3.24}$$

where $w_{z1} = \frac{1}{R_2 C_1}$ and $w_{p3} = \frac{1+\frac{R_2}{R_1}}{R_2 C_1}$. In our design, w_{p3} is 4 times bigger than the zero, since the resistive ratio is 4. This extra zero will improve the phase margin of the LDO.

3.3.3 Experimental Results

The proposed circuit was fabricated in a standard 0.35μm CMOS process. The supply voltage of the proposed circuit is $2.5V$. The measurement results of both versions of this charge-pump will be presented in the rest of this section. In this section, the first version charge pump refers to the charge pump with a $3.14pF$ pumping capacitor and the second version charge pump refers to the charge pump with a $4.72pF$ pumping capacitor.

The measured transfer curves from V_{targ} to V_{out} is shown in Fig. 3.24. The steady-state output voltage is regulated to $5V_{targ}$. Deviation at the low voltages in Fig. 3.24(a) and 3.24(b) is caused by the error-amplification OTAs input transistors in the LDO being pushed out of the saturation region. Deviation at the high voltages is caused by the limited output voltage of the open loop charge-pump. The second version can provide higher output voltage because the term $N\frac{I_L}{C_P f}$ in

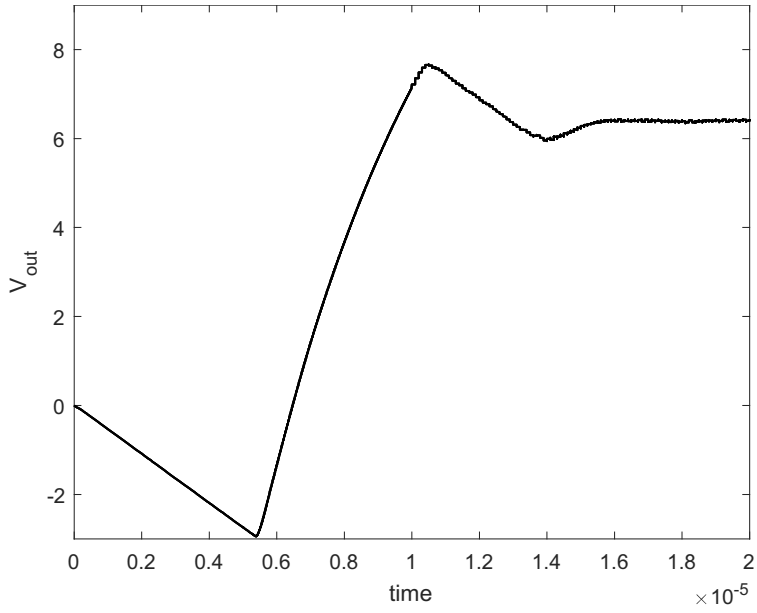


Figure 3.21: Output of the LDO (V_{out}) presented in Fig. 3.19

equation (3.1) is a smaller term in the second version.

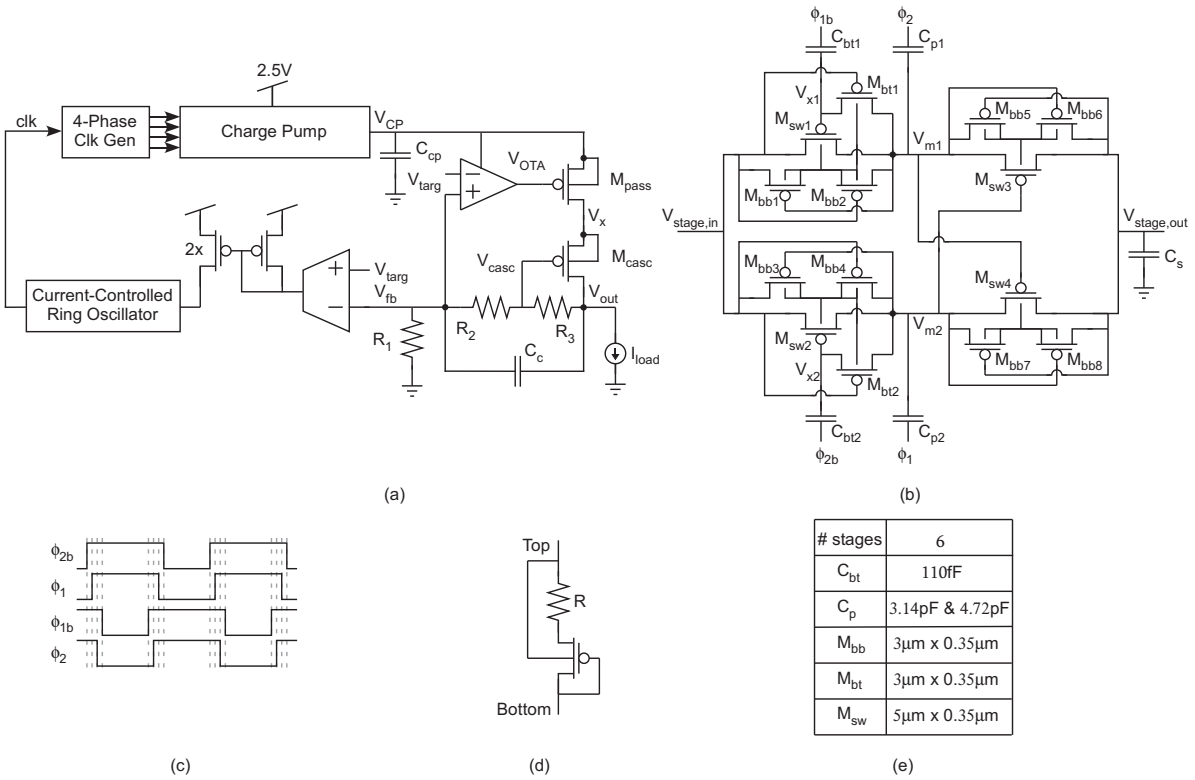
The load-regulation measurements of the proposed circuit are shown in Fig. 3.25. These plots show that the second version of this charge-pump handles higher load currents. Fig. 3.26 shows the measured output resistance of this charge-pump for both versions. To measure these two plots the slope of the curves between 40μA and 50μA in Fig. 3.25 was calculated. These two plots show that the output resistances of both versions are close. This is because both structures use the same exact LDO structures with exact same transistor sizes. However, the loop gain of these two structures are slightly different due to the different pumping capacitor sizes. Thus, there is a slight difference between the closed-loop output impedance of these two charge-pumps. Fig. 3.27 shows the measured output resistance of the proposed charge pump as a function of G_m. In order to take this measurement, the bias current of the OTA inside the charge-pump loop was set to 2μA, 4μA, and 5μA.

Fig. 3.28 shows the transient response of both charge pumps under 40μA of load current with no load capacitor in the output. The Enable signal activates the charge pump for 3.5ms. The rise times of the charge pumps are less than 0.5ms which are fast enough for the injection procedure.

The efficiency of a charge pump can be calculated by:

$$\gamma = \frac{V_{out} I_{load}}{V_{dd} I_{vdd}} \tag{3.25}$$

which is the power delivered to the load divided by the total power going into the charge-pump circuit. Fig. 3.29 and 3.30 show the measured efficiency of the proposed charge pump. A 6.5V pulse is typically used to inject the FG transistors. Fig. 3.29 shows that the maximum efficiency for a 6.5V output occurs at 70μA of load current. The maximum efficiency at 6.5V is around 20%.



# stages	6
C_{bt}	110fF
C_p	3.14pF & 4.72pF
M_{bb}	3µm x 0.35µm
M_{bt}	3µm x 0.35µm
M_{sw}	5µm x 0.35µm

Figure 3.22: Circuit structure of the proposed injection charge-pump

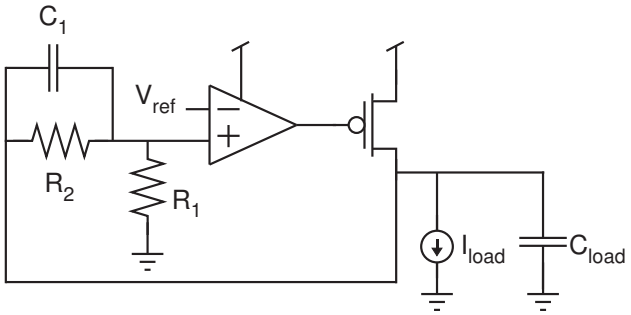


Figure 3.23: Circuit structure of the feed-forward compensation technique

Since the charge-pump is activated in a short period of time, a low-efficiency number is not a severe issue in FG programming.

3.4 Conclusion

In this Chapter, three high-voltage charge pumps are presented that are suitable for tunneling and injection of FG transistors. We have presented the design and results of a regulated high-voltage charge pump for generating tunneling voltages to program FG transistors. This compact charge pump leverages variable-frequency regulation and a new circuit for minimizing short-circuit current to provide reliable tunneling voltages while consuming very little power. Because of its compact size and low-power operation, this charge pump is ideally suited for battery-powered applications.

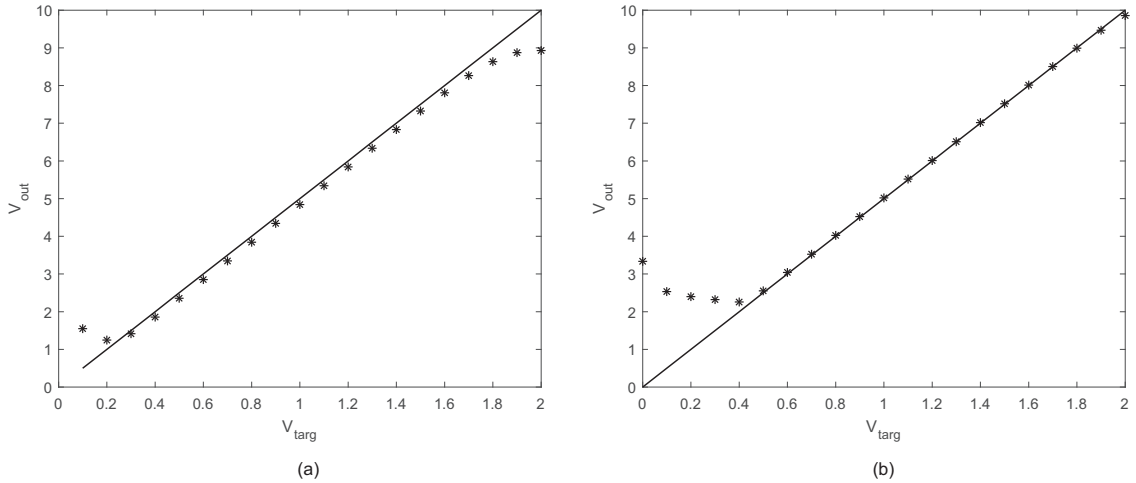


Figure 3.24: Measured transfer curve (a) first version charge-pump (b) second version charge-pump.

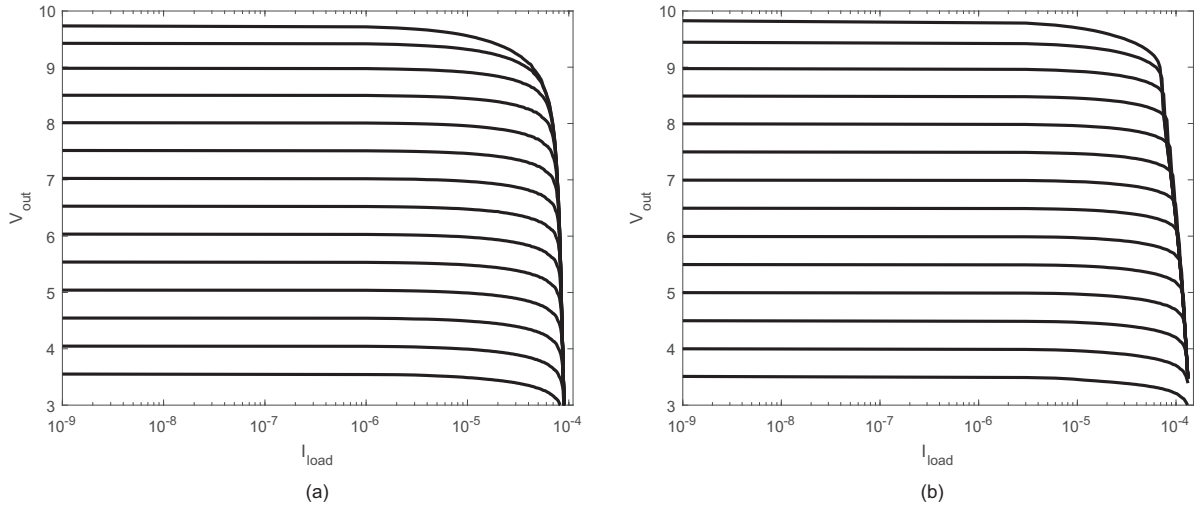


Figure 3.25: Measured load regulation of (a) first version charge-pump (b) second version charge-pump.

Additionally, it can easily be adapted to provide higher or lower voltages for programming FG transistors in scaling CMOS processes and also for other applications, such as MEMS devices.

Additionally, the design of two regulated charge pumps for injection of FG transistors are presented. A high PSRR LDO is used inside the charge pump loops to reduce the ripple. A feed-forward compensation technique is used in the LDO to improve the phase margin of the LDO. This is very critical in FG injection because an under-damped voltage is not suitable for FG injection. The proposed charge pumps have compact sizes and consume low power which makes them suitable for battery-powered applications.

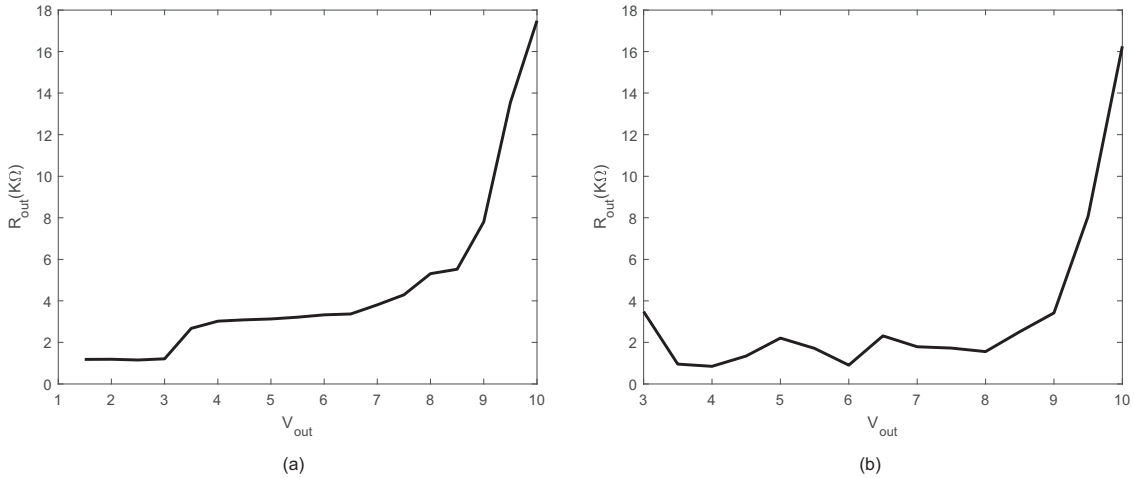


Figure 3.26: Measured output resistance of (a) first version charge-pump (b) second version charge-pump.

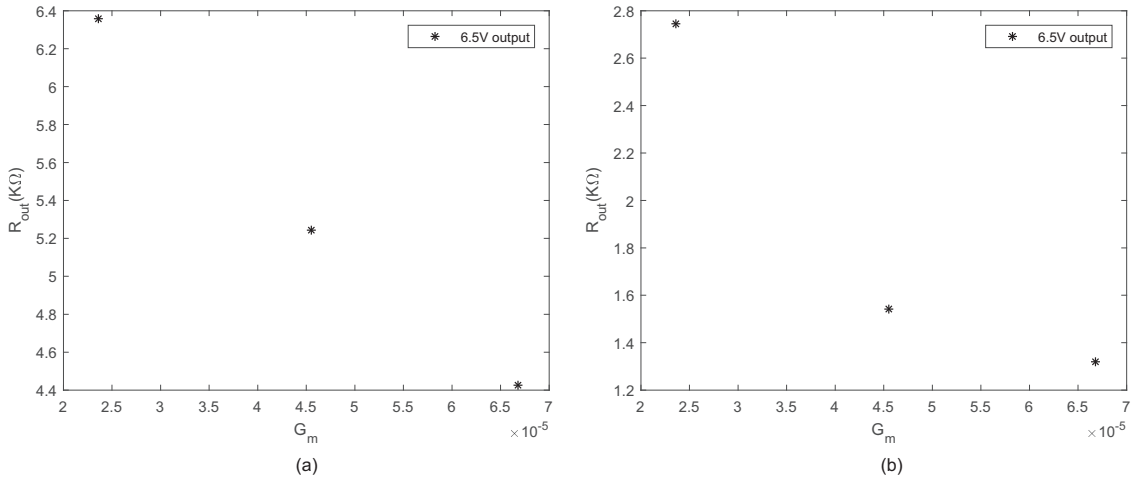


Figure 3.27: Measured output impedance as a function of Gm for (a) first version charge-pump (b) second version charge-pump.

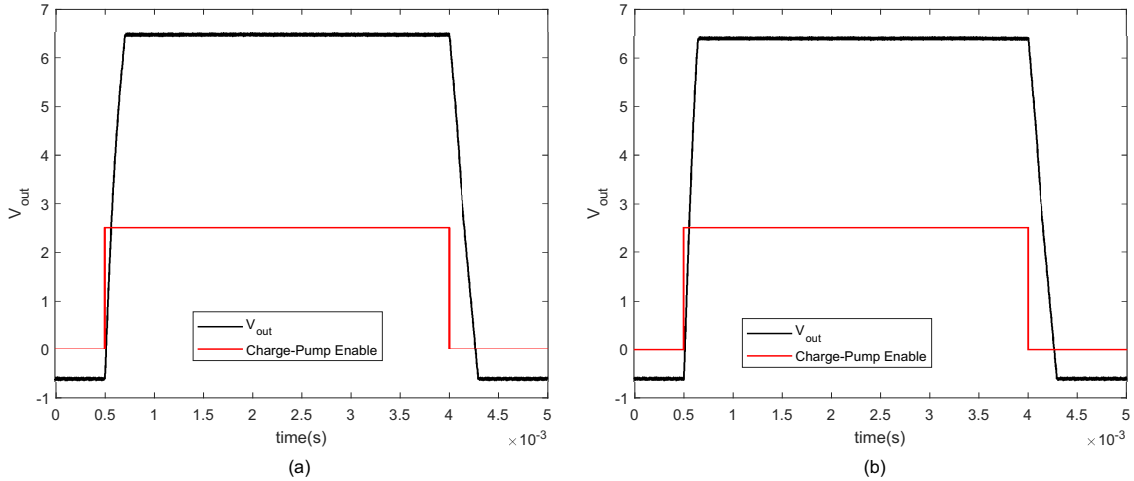


Figure 3.28: Measured transient response of the proposed closed loop charge-pump under 40μA load and no load capacitance for (a) first version charge-pump (b) second version charge-pump.

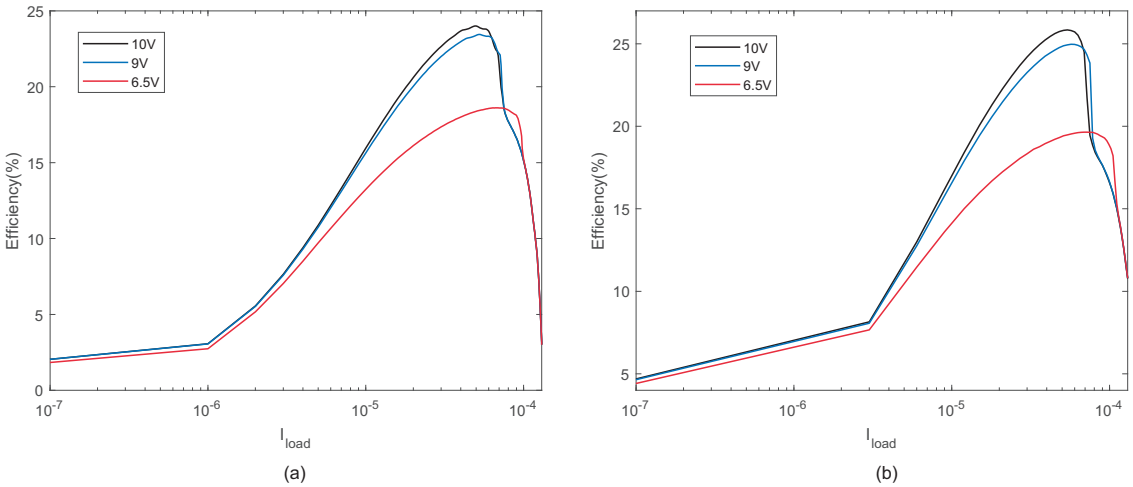


Figure 3.29: Measured efficiency of the proposed charge pump vs. I_{load} for (a) first version charge-pump (b) second version charge-pump.

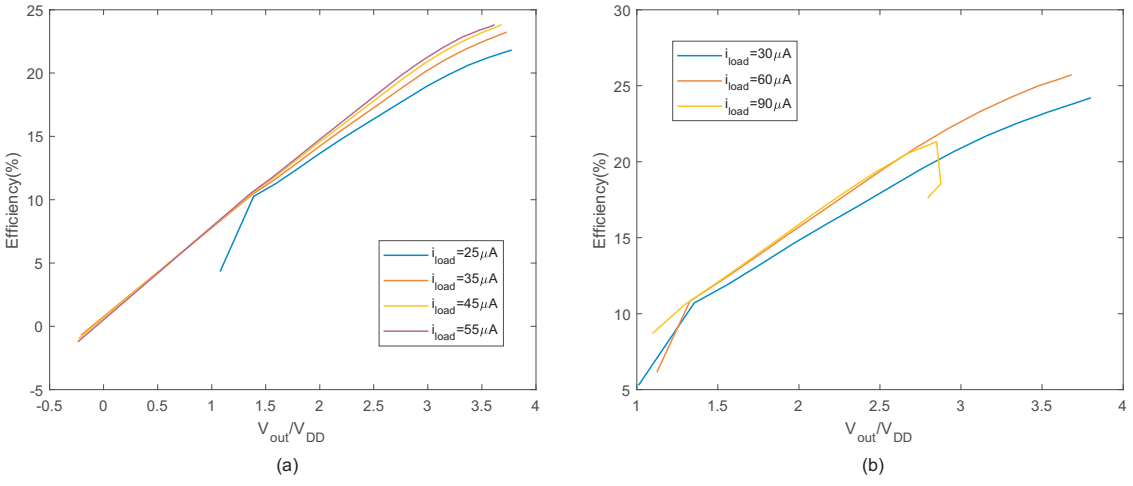


Figure 3.30: Measured efficiency of the proposed charge pump vs. V_{out}/V_{DD} for (a) first version charge-pump (b) second version charge-pump.

Chapter 4

A Low-Power Voltage Reference Cell with Above-$1V$ Output

A low-power voltage reference cell for system-on-a-chip applications is presented in this chapter. The proposed cell uses a combination of standard transistors and thick-oxide transistors to generate a voltage above 1V. A design procedure is also presented for minimizing the temperature coefficient (TC) of the reference voltage. This circuit was fabricated in a standard 0.35μm CMOS process. It generates a 1.52V output with a TC of 42ppm/°C from -70°C to 85°C while consuming only 1.11μW.

4.1 Background of the voltage reference cells

Voltage reference circuits are critical building blocks that are used to generate a stable voltage across a wide range of temperatures. As newer CMOS technology nodes are providing lower threshold voltages and reduced supply voltages, the main emphasis of recent designs has been to generate a low reference voltage (especially less than 1V) using a low supply voltage. Many good designs have been developed that provide these sub-1V reference voltages using very low power.

However, many applications still exist that require reference voltages greater than 1V, but also consume very little power. While a number of good above-1V designs have been presented, there has been a severe tradeoff between having a low temperature coefficient (TC) and low power consumption.

In this chapter, we present a voltage reference cell that is able to simultaneously achieve a low TC and low power consumption over a large range of temperatures. Our voltage reference cell has been fabricated in a standard 0.35μm CMOS process and is capable of generating a reference voltage greater than 1V with a TC<110ppm/°C and single μW power consumption over a wide range of temperatures. To achieve this performance, we use the difference between the threshold voltages of standard (thin-oxide) transistors and I/O thick-oxide transistors to provide both elevated voltages

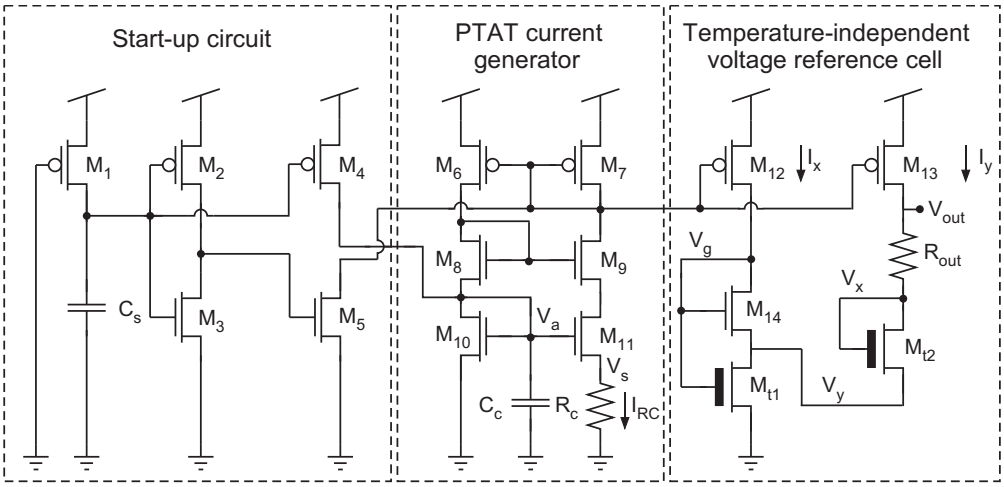


Figure 4.1: Schematic diagram of the proposed voltage reference cell. M_{t1} and M_{t2} are thick-oxide transistors (5V I/O devices).

and also the temperature characteristics needed to minimize the TC. The generated reference voltage is a function of the ratio of resistors, and thus, the temperature-dependent terms due to the resistors largely cancel out.

4.2 Proposed Voltage Reference Cell

Figure 4.1 illustrates the proposed voltage reference cell. All the transistors used in this circuit are standard (thin-oxide) transistors except for M_{t1} and M_{t2} which are thick-oxide transistors (5V I/O devices). Two resistors are used in this circuit—one to generate the reference current (R_C) and the other to generate the final reference voltage (R_{out}).

This circuit has three main blocks, as shown in Fig. 4.1. A current reference cell is used to generate a current that is proportional to absolute temperature (PTAT). A start-up circuit is used to initialize the current reference cell and stabilize its current at a nonzero value. Finally, the third block generates the temperature-independent voltage by first creating a voltage at node V_x that is complementary to absolute temperature (CTAT). Resistor R_{out} then combines the PTAT and CTAT signals into an overall output voltage that has very low temperature dependence.

To achieve low power consumption, our circuit was designed to operate in the subthreshold region with very low bias currents. The drain current of a transistor biased in subthreshold can be expressed as [57]:

$$I_d = I_0' S e^{\frac{\kappa(V_g - V_T)}{U_T}} \left(e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}} \right) \tag{4.1}$$

where $I_0' = 2\eta\mu C_{ox} U_T^2$, S is the aspect ratio (W/L) of the transistor, $U_T = kT/q$ is the thermal voltage, k is the Boltzmann constant, q is the elementary charge, T is the absolute temperature in kelvins, V_T is the MOSFET threshold voltage, and $\eta = 1/\kappa$ is the subthreshold slope. Voltages

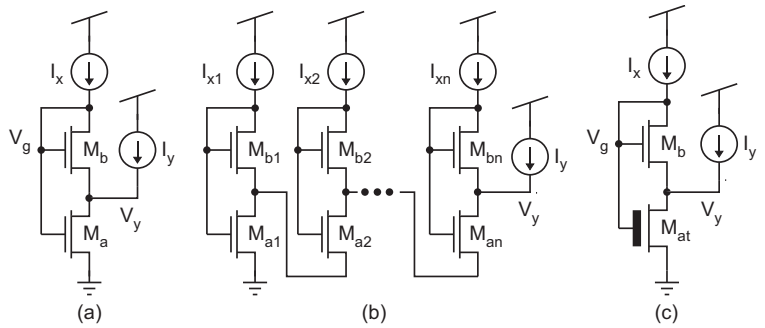


Figure 4.2: (a) A PTAT voltage generator (b) Generating a higher voltage through stacking PTAT voltage generators (C) A CTAT voltage generator.

V_g, V_s, and V_d are the gate, source, and drain voltages, respectively, referenced to the substrate. When the transistor operates in the saturation region, $exp(\frac{-V_d}{U_T})$ approaches zero and is, therefore, negligible. We also assume that the transistors have been designed with large-enough channel lengths that the channel-length modulation effect can be safely neglected.

The detailed operation of each block will be explained in the remainder of this section.

4.2.1 Reference Current Generator

We use a standard PTAT current generation block which includes a current source and a current mirror. The difference between the V_{gs} values of M_{10} and M_{11} establishes a voltage across R_C. By proper sizing of R_C, M_{10}, and M_{11} ($S_{11} > S_{10}$), all the transistors in the PTAT current generator can be biased to be in the subthreshold region. By setting $S_6 = S_7$, the PMOS current mirror provides unity gain. Using (4.1) to define expressions for the currents in M_{10} and M_{11}, the current generated by R_C is

$$I_{R_C} = \frac{V_s}{R_C} = U_T \left(\frac{1}{R_C}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{4.2}$$

which linearly increases with temperature. The off-chip capacitor (C_C) is used to serve as a low-pass filter to stabilize the gate voltage of M_{10} and M_{11}. I_{R_C} is mirrored to M_{12} and M_{13} with ratios of x and y. Thus, we have

$$I_x = I_{12} = I_{14} = xI_{R_C} \tag{4.3}$$

$$I_y = I_{13} = I_{t2} = yI_{R_C} \tag{4.4}$$

Accordingly, the voltage drop across R_{out} caused by the PTAT current source can be expressed as

$$V_{PTAT} = y\left(\frac{kT}{q}\right)\left(\frac{R_{out}}{R_C}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{4.5}$$

This voltage is PTAT and can be set by R_{out}/R_C and the aspect ratios for M_{10} and M_{11}

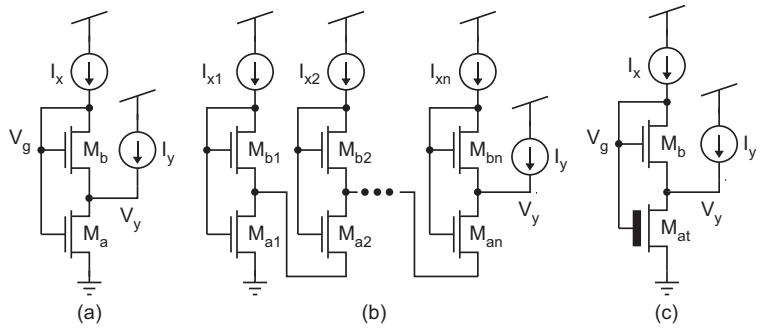


Figure 4.3: (a) A PTAT voltage generator (b) Generating a higher voltage through stacking PTAT voltage generators (C) A CTAT voltage generator.

4.2.2 Start-Up Circuit

A start-up circuit is used to initialize the current reference cell and stabilize its current at a nonzero value ($I_{R_C} \neq 0$). We use the start-up circuit presented by [58]. The start-up circuit turns on M_6, M_7, M_{10}, and M_{11} at power-up to initialize a nonzero I_{R_C}. Afterwards, M_4 and M_5 turn off and the entire start-up circuit no longer draws current.

4.2.3 Temperature-Independent Voltage Reference Cell

The temperature-independent voltage reference cell is based upon a modified version of the building block shown in Fig. 4.3(a) that is commonly used as a PTAT voltage generator (e.g. [59]). The currents through M_b and M_a are I_x and $I_x + I_y$, respectively. These two currents can be expressed using (4.1) as

$$M_b : I_x = I'_0 S_b e^{\frac{\kappa(V_g - V_T)}{U_T}} e^{-\frac{V_y}{U_T}} \tag{4.6}$$

$$M_a : I_x + I_y = I'_0 S_a e^{\frac{\kappa(V_g - V_T)}{U_T}} \left(1 - e^{-\frac{V_y}{U_T}}\right) \tag{4.7}$$

By dividing (4.7) by (4.6) and solving for V_y, we find that

$$V_y = U_T \ln\left(1 + \left(\frac{S_b}{S_a}\right)\left(\frac{I_x + I_y}{I_x}\right)\right) \tag{4.8}$$

The resulting voltage is PTAT and can be set by proper sizing of M_a and M_b and the bias currents (I_x and I_y). However, the value of V_y cannot be made large because of the log compression working on the ratio of the transistor sizes and currents. Some designs stack this circuit repeatedly to generate a higher reference voltage, as shown in Fig. 4.3(b), but they still struggle to achieve a high voltage [60].

Figure 4.3(c) shows a modification to Fig. 4.3(a) that replaces the bottom transistor with a thick-oxide device. This configuration has been used previously to provide a low voltage at V_y that is CTAT [58, 61, 62]. We, instead, use this configuration to help us achieve a much higher voltage

at another node that is also CTAT.

The currents through M_b and M_{at} follow the forms of (4.6) and (4.7), with the only differences being the various process-dependent parameters for the thick-oxide device, M_{at},

$$M_{at} : I_x + I_y = I'_{0t} S_{at} e^{\frac{\kappa_t (V_g - V_{Tt})}{U_T}} \left(1 - e^{-\frac{V_y}{U_T}}\right) \tag{4.9}$$

where the extra 't' in the subscript represents the value for the thick-oxide device. By dividing (4.9) by (4.6) and solving for V_y, we obtain

$$V_y = U_T \ln \left(1 + \frac{I_x + I_y}{I_x} \frac{I'_0 S_b}{I'_{0t} S_{at}} e^{\frac{\kappa_T V_{Tt} - \kappa V_T}{U_T}} e^{\frac{\kappa V_g - \kappa_t V_g}{U_T}}\right) \tag{4.10}$$

Noting that $V_{Tt} > V_T$ and that $\kappa V_g - \kappa_T V_g \approx 0$, (4.10) can be approximated as

$$V_y \approx \kappa_t V_{Tt} - \kappa V_T + U_T \ln \left[\left(\frac{I_x + I_y}{I_x}\right)\left(\frac{I'_0 S_b}{I'_{0t} S_{at}}\right)\right] \tag{4.11}$$

with proper sizing of the transistors. V_{Tt} is typically several hundred mV greater than V_T, which means that using a thick-oxide transistor for M_{at} can produce a voltage at V_y that is much larger that can be produced by Fig. 4.3(a). Additionally, V_y is dominated by the first two terms of (4.11), and since threshold voltages are widely known to be CTAT [63], V_y is also a CTAT voltage.

Next, we show that we can use Fig. 4.3(c) to create a CTAT voltage at the bottom of R_{out} (V_x) in Fig. 4.1. The current through M_{t1} can be expressed as

$$I_{t1} = \left(1 + \frac{y}{x}\right) I_{14} \tag{4.12}$$

Using (4.11) and (4.12), the following equation can be achieved

$$V_y \approx \kappa_t V_{Tt} - \kappa V_T + U_T \ln \left[\left(1 + \frac{y}{x}\right)\left(\frac{I'_0 S_{14}}{I'_{0t} S_{t1}}\right)\right] \tag{4.13}$$

Using (4.13) and the equation for the current through M_{t2}, the voltage at node V_x can be expressed as

$$V_x = 2V_{Tt} - \frac{\eta_t}{\eta} V_T + \eta_t U_T \ln \left[\frac{\xi}{R_c T}\right] \tag{4.14}$$

where

$$\xi = \left(y + \frac{y^2}{x}\right) \frac{S_{14}}{S_{t1} S_{t2}} \ln \left(\frac{S_{11}}{S_{10}}\right) \frac{q}{k} \frac{\eta C_{ox}}{2\eta_t^2 \mu C_{oxt}^2} \tag{4.15}$$

V_x is dominated by the first two terms of (4.14), and since threshold voltages are widely known to be CTAT [63], V_x is a CTAT voltage.

Finally, V_{out} is the summation of the PTAT voltage given by (4.5) and the CTAT voltage given

by (4.14).

$$V_{out} = yR_{out}I_{R_C} + V_x \tag{4.16}$$

$$V_{out} = y\frac{R_{out}}{R_C}U_T \ln\left(\frac{S_{11}}{S_{10}}\right) + 2V_{Tt} - \frac{\eta_t}{\eta}V_T + \eta_t U_T \ln\left[\frac{\xi}{R_c T}\right] \tag{4.17}$$

The goal of this voltage reference cell is to generate an above-$1V$ output using thick-oxide transistors. The execution of idea is visible in equation (4.17) where $2V_{Tt}$ is a considerable portion of the output voltage. Additionally, the voltage drop across R_{out} further increases V_{out}. Thus, using the proposed structure, a higher voltage can be achieved by using small number of transistors and taking advantage of the thick-oxide transistor's larger threshold voltage.

4.3 Design procedure for low TC V_{out}

In this section, we discuss how to design this voltage reference circuit to have a low temperature coefficient. By taking the derivative of (4.17) with respect to T, we arrive at an equation for the TC at V_{out}.

$$TC = \frac{\partial V_{out}}{\partial T} = y\frac{k}{q}\frac{R_{out}}{R_C}\ln\left(\frac{S_{11}}{S_{10}}\right) + 2\alpha_t - \frac{\eta_t}{\eta}\alpha + \frac{\eta_t k}{q}\left[\ln\left(\frac{\xi}{R_c T}\right) - 1\right] \tag{4.18}$$

The α terms, which have a negative value, come from the commonly used expression for the temperature effects on the threshold voltage [63] given by

$$V_T(T) = V_T(T_0) + \alpha(T - T_0) \tag{4.19}$$

where T_0 is a reference temperature, and T is the temperature of interest.

Analyzing (4.18), we can see that the temperature dependence of the resistors has little impact on the TC; assuming both resistors are made from the same material, their temperature dependencies cancel in the first term. R_C is also contained within the ln term, but its temperature effects have little impact since they are compressed by the logarithmic function; hence, the temperature effects of R_C are neglected in this analysis (and they are far less significant than the T term in the same ln function). The ξ term also contains the temperature-dependent items μ, η, and η_t; again, since they are logarithmically compressed, they have little impact on the overall TC, and their temperature effects can be safely neglected.

The only other remaining temperature-dependent term in (4.18) is T within $\ln(\xi/R_C T)$. Since T is within an ln function, its affect on the TC is significantly compressed, meaning that it is possible to achieve a low overall TC. To achieve a low TC over a temperature range of interest, a reference temperature in the middle of the range should be chosen; then, the TC can be minimized

by setting (4.18) equal to zero at that reference temperature, T_0. By then solving for $\ln(\xi/R_C T)$ and plugging in that expression into (4.17), we find the optimal value for V_{out}.

$$V_{out} = 2V_{Tt} - \frac{\eta_t}{\eta}V_T - 2\alpha_t(T_0) + \frac{\eta_t}{\eta}\alpha(T_0) + \eta_t U_T \tag{4.20}$$

$$\approx 2V_{Tt} - V_T - 2\alpha_t(T_0) + \alpha(T_0) + \eta_t U_T \tag{4.21}$$

Noting that the α terms have negative values (i.e. V_T is CTAT) and are typically in the mV/K range [63], the optimal V_{out} at room temperature will be greater than 1V, particularly because of the reasonably high V_T of 5V I/O devices.

Once again, we note that the transistor aspect ratios in (4.17) are primarily within ln functions, so the ratio of R_{out}/R_C plays an important factor in establishing the correct V_{out}. To determine the best ratio for R_{out}/R_C to minimize the TC, (4.17) and (4.20) can be equated, and the best resistor ratio can be found to be

$$y\frac{R_{out}}{R_C} = \frac{\eta_t\left[1 - \ln\left(\frac{\xi}{R_C T_0}\right)\right] - \frac{q}{k}(2\alpha_t - \alpha)}{\ln\left(\frac{S_{11}}{S_{10}}\right)} \tag{4.22}$$

In our design, a ratio of $R_{out}/R_C \approx 6$ is needed to achieve a good TC. Additionally, by changing the resistive ratio slightly, the temperature at which the TC is minimized can be shifted to a higher or lower value while still maintaining a good TC; varying the resistor ratio is explored further in section 4.4.

Next, we present a design procedure to obtain a low-TC, above-1V voltage reference cell using the circuit of Fig. 4.1.

1. Set $\frac{S_{11}}{S_{10}} \gg 1$, and choose a proper size for R_C to bias the current reference cell in subthreshold. Note that choosing a large $\frac{S_{11}}{S_{10}}$ will result in a larger R_C to bias the circuit in subthreshold, so $\frac{S_{11}}{S_{10}}$ must not be too large.

2. Choose appropriate transistor aspect ratios such that $\frac{S_{14}}{S_{t1}S_{t2}} \geq 1$ and also that x and y keep all transistors in subthreshold. These design choices make it possible to approximate (4.10) with (4.11).

3. Ensure that all transistor lengths are large enough to neglect the effects of channel-length modulation.

4. Choose the midpoint, T_0, of the desired temperature range, and use (4.22) to solve for the ratio $y\frac{R_{out}}{R_C}$.

5. If the midpoint of the V_{out} vs. T curve is not at the desired location, then according to (4.22), we can adjust R_C to move T_0 to higher or lower temperatures. Also, R_{out} must be adjusted,

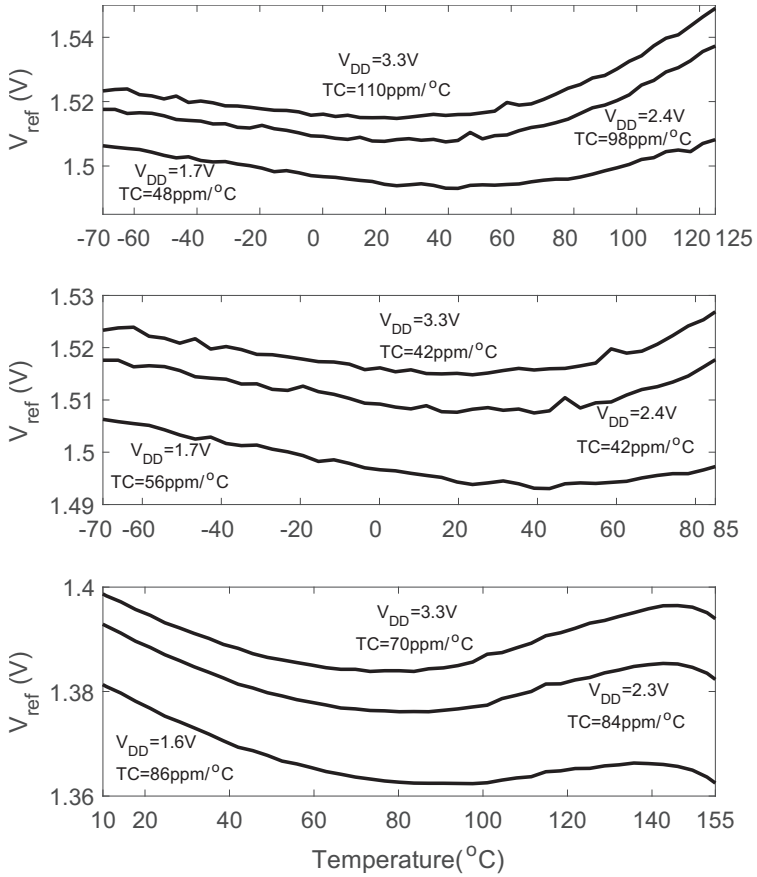


Figure 4.4: Measured TC of the proposed circuit under three different supply values for $R_{out}/R_C = 3M\Omega/500k\Omega$ (top and middle) and $R_{out}/R_C = 2M\Omega/400k\Omega$ (bottom).

accordingly, to keep the resistive ratio at a reasonable value, based on (4.22).

4.4 Experimental Results

The voltage reference circuit of Fig. 4.1 was fabricated using a standard $0.35\mu m$ CMOS technology. The die area of the active portion of this circuit was $0.033mm^2$, and the supply voltage of this circuit was 3.3V. The resistors (R_C and R_{out}) used in this circuit were off-chip resistors to allow for variety of R_{out}/R_C combinations. Using the design procedure of section 4.3, we found that $R_{out}/R_C \approx 6$ for a low TC output. R_{out} and R_C were set to 3MΩ and 500kΩ, respectively, to achieve this ratio and to ensure subthreshold operation. These values were also chosen to minimize the TC around room temperature.

Figure 4.4(a)&(b) shows the measured V_{out} over temperature for multiple values of V_{DD}. Figure 4.4(b) shows that this circuit is able to achieve a low TC of 42ppm/°C over the temperature range of interest (-70 to +85°C). Even extending this temperature range significantly to cover from -70 to +125°C, this circuit can still achieve a TC of 110ppm/°C at the nominal V_{DD}=3.3V (see Fig. 4.4(a)). By reducing V_{DD} to 1.7V, the TC is 48ppm/°C over this extended temperature range.

As discussed in the previous section, this circuit can be optimized to achieve a good TC for

Table 4.1: Comparison of the proposed work with other works

	Process	TC (ppm/°C)	T-range	V_{REF}	LR (mV/V)	PSRR (dB)	V_{DD}	Power	Area (mm^2)	Comments
[59]	0.35μm	215-394	[-20 80]	1.18V	4.5	-	1.3∼3.3	0.108μW	0.21	-
[64]	0.18μm	73	[0 100]	1.25V	0.31	-41dB@100Hz	1.4∼3.6	35pW	0.0025	Native NFETs
[65]	0.18μm	4.1	[-55 125]	1.1402V	0.3	-54dB@100Hz	1.3∼2.6	11.18μW	0.05	NPN BJTs
[66]	0.6μm	14.36	[0 100]	1.2525V	5.5	-42dB@10MHz	1.5∼2	40μW	0.11	-
[67]	0.25μm	627	[20 50]	1.03V	0.2	-51dB@100Hz	1.5∼3.5	0.12μW	0.011	-
[68]	0.18μm	147	[-40 120]	1.09V	-	-62dB@100Hz	1.2∼1.8	0.1μW	0.0294	-
[69]	0.18μm	4	[0 100]	1.012V	0.5	-66dB@1kHz	1.1∼1.8	21μW	-	-
[70]	0.35μm	12.85	[5 95]	1.2V	28	-26.2dB@100Hz	1.75∼3.5	35.7μW	0.0206	-
This work #1	0.35μm	110@3.3V	[-70 125]	1.52V	10	-44dB@10kHz	1.7∼3.3	1.11μW	0.033	-
This work #1	0.35μm	42@3.3V	[-70 85]	1.52V	10	-44dB@10kHz	1.7∼3.3	1.11μW	0.033	-
This work #2	0.35μm	70@3.3V	[10 160]	1.395V	9.33	-44.8dB@10kHz	1.6∼3.3	1.34μW	0.033	-

different temperature ranges. Figure 4.4(c) shows an example of shifting the midpoint of the temperature range to a higher value. This midpoint reference temperature can be increased by setting R_{out}/R_C to a smaller value and by also reducing the value of R_C, as is evidenced by (4.18). In this example, R_C was reduced to 400kΩ which increases the temperature midpoint, T_0, due to the $\ln(\xi/R_C T)$ term. R_{out} was also decreased to 2MΩ to keep the resistor ratio similar (but slightly smaller). The result is that the same circuit, with different resistor values, can be used to provide a low TC (70ppm/°C at V_{DD}=3.3V) at higher temperatures (+10 to +155°C).

The measured supply current versus temperature for these two reference voltages (V_{out}=1.52V, 1.395V) under a 3.3V supply is shown in Fig. 4.5. The power consumption of this circuit at room temperature for the two R_{out}/R_C conditions were 1.11μW (V_{out}=1.52V) and 1.34μW (V_{out}=1.395V). Thus, this circuit is an appropriate choice for low-power applications.

The line regulation of the circuit was measured at room temperature by sweeping the supply voltage from 0V to 3.3V, as shown in Fig. 4.6. The line regulation for the two R_{out}/R_C conditions of Fig. 4.4 were 10mV/V (V_{out}=1.52V) and 9.3mV/V (V_{out}=1.395V). The line regulation can be significantly improved by cascoding the PMOS transistors M_6, M_7, M_{12}, and M_{13}. The limited $V_{DD,min}$ in these two cases was due to the high output voltage of the circuit. In both cases, $V_{DD,min}$ was approximately $V_{DD,min} \approx V_{out} + 200$mV. Additionally, the measured power supply rejection ratios (PSRR) at 100Hz were -35dB (V_{out}=1.52V) and -37.9dB (V_{out}=1.395V). The PSRR values at 1MHz were -44dB (V_{out}=1.52V) and -44.8dB (V_{out}=1.395V).

The die-to-die distribution of the DC output voltage was found at room temperature by measuring the average reference voltage (μ) and the standard deviation (σ) for 17 available chips. The coefficient of variation (σ/μ) for the 1.52V and 1.395V outputs were 2% and 1.76%, respectively.

Table 4.1 compares the three cases of our voltage reference circuit shown in Fig. 4.4 to other similar circuits. Specifically, we compared our work to other circuits that 1) were fabricated, 2) were in a CMOS process, 3) have $V_{out}\geq$1V, and 4) have power consumption $\leq$50μW. As can be seen from this Table, our voltage reference circuit is able to provide a good balance of a low TC, low

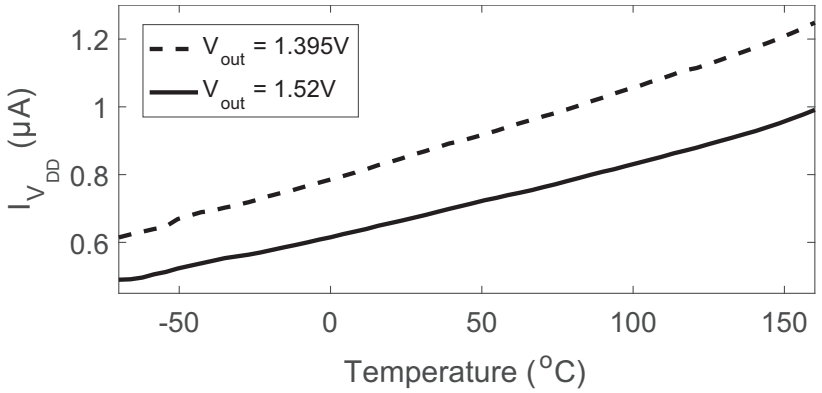


Figure 4.5: Measured supply current.

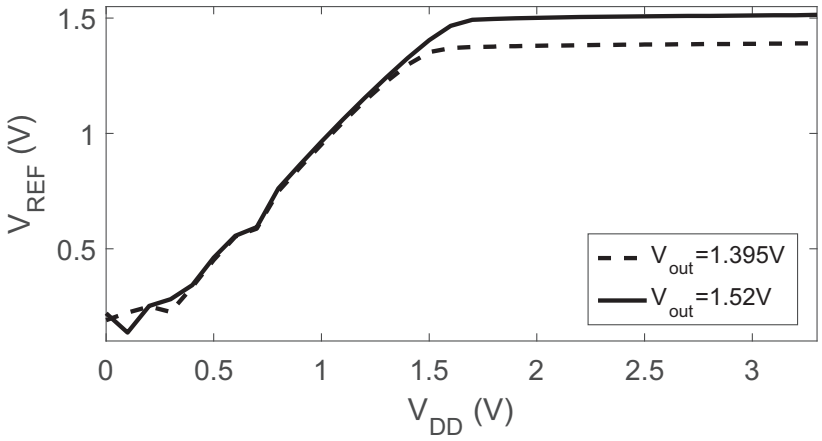


Figure 4.6: Measured line regulation.

power consumption, and a large range of temperatures. Additionally, our circuit only uses devices available in standard CMOS processes (thicker-oxide I/O devices are now widely available), whereas some of the listed designs require non-standard devices [64, 65]. Improving the line regulation of our circuit can easily be accomplished by cascoded structures at the expense of a higher minimum supply voltage.

4.5 Conclusion

A low power voltage reference cell for system-on-a-chip applications has been presented in this chapter. This proposed cell uses a combination of thin-oxide and thick-oxide transistors to generate a reference voltage greater than 1V with a low TC. We also presented a design methodology for how to translate this circuit to other processes and to provide a low-power and low-TC reference voltage.

Chapter 5

On-chip Voltage Regulators for Supply and Mid-Rail Voltage References

Every electronic circuit includes reference voltages (e.g. analog/digital supply voltage, mid-rail voltage). These reference voltages must be constant over a wide range of load currents. Also, they must have low sensitivity to the input battery voltage drop or fluctuations (e.g. battery droop, environmental noises). A voltage regulator is used to provide this constant DC output voltage and to continuously keep the output voltage at the desired value regardless of changes in load current or supply voltage. In this work, we need a voltage regulator for both the analog supply voltage (AV_{dd}) and the digital supply voltage (DV_{dd}). Also, in some cases, the supply voltage must be able to provide both sinking and sourcing load currents. As an example, the mid-rail voltage regulator is a voltage regulator that must provide a voltage that is half the supply voltage ($V_{dd}/2$). In addition to its sourcing capability, this mid-rail voltage regulator must be able to sink some load current and hold the output voltage constant at $V_{dd}/2$. In the remainder of this Chapter, I will discuss the voltage regulators that I have designed to be used for the on-chip programming of the FG transistors and the systems using these FG transistors.

5.1 Quick Review of Voltage Regulators

A simple linear voltage regulator is a voltage-controlled current source used to force a fixed voltage to the output terminal of the voltage regulator. The control circuit senses the fluctuation at the output terminal and adjusts the current source based on the load current to hold the output voltage at the desired value. Fig. 5.1 shows a simplified model of a typical voltage regulator. The design limit of the current source defines the maximum load current that the voltage regulator can provide and still maintain the regulation. In general, a simple CMOS voltage-controlled current

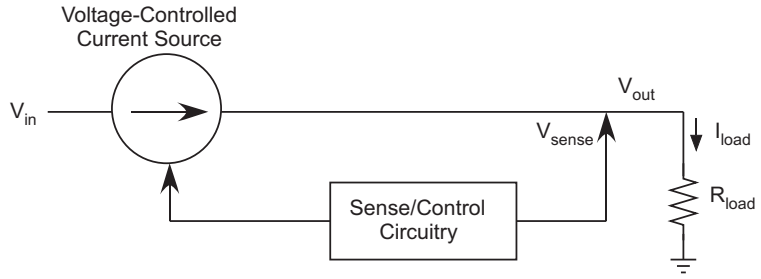


Figure 5.1: A simplified model of a voltage regulator.

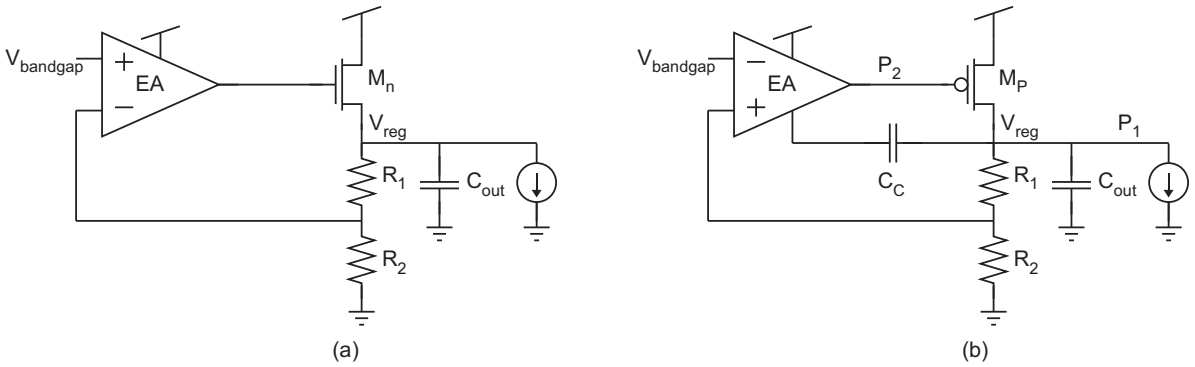


Figure 5.2: Two conventional voltage regulators.

source is realized using either a pMOS transistor or a nMOS transistor. These two structures are shown in Fig. 5.2. The pass device (M_p or M_n) in this voltage regulator is realized by a pMOS or an nMOS transistor. A regulator using an nMOS transistor is a standard voltage regulator. A regulator using a pMOS transistor is called a low drop-out regulator (LDO). The current flowing out of the drain terminal of the pass transistor is controlled by M_p or M_n and the voltage amplifier. This current is largely the load current. The current through the resistive voltage divider is assumed to be negligible (R_1 and R_2) compared to the load current. The feedback loop, which controls the output voltage, is obtained by using R_1 and R_2 to "sense" the output voltage and applying the sensed voltage to the inverting input of the voltage amplifier. The non-inverting input of the amplifier is tied to a voltage reference. The output voltage of the amplifier is connected to the gate of the pass transistors, which means the circuit is continuously providing an output voltage and a constant load current through the pass transistor to force the voltages at its inputs to be equal. The feedback loop acts continuously and holds the regulated output at $(1 + \frac{R_2}{R_1})V_{out}$ regardless of the changes in the load current. A sudden increase or decrease in the load current will cause the output voltage to change until the loop can correct and stabilize to the new level, which is called the transient response. The increase or decrease in the output voltage is sensed through R_1 and R_2 and appears at the output of the error amplifier, which results in the correction of the output voltage through I_{M_p} or I_{M_n}. In the next few sections, I will review some basic characteristics of voltage regulators.

A. Dropout Voltage

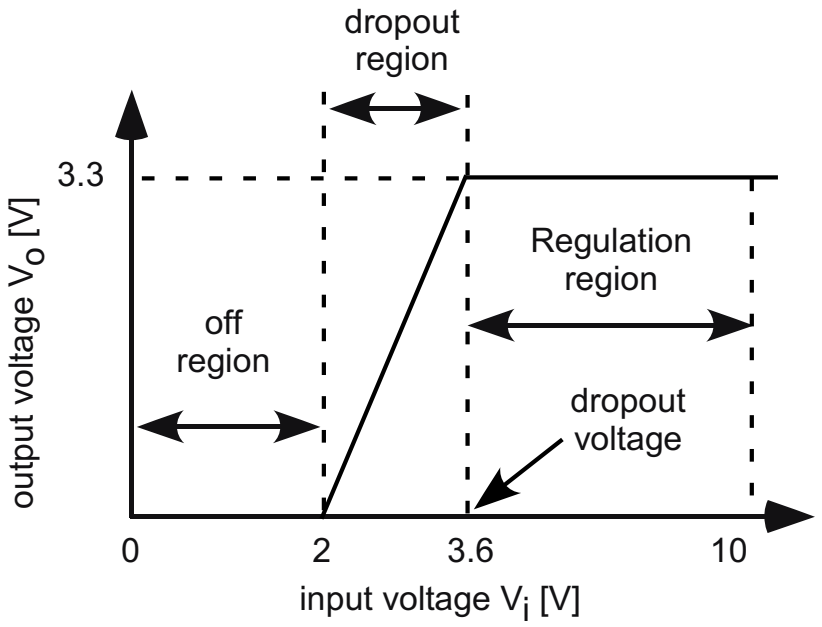


Figure 5.3: Dropout region of an LDO regulator

The input-to-output differential voltage at which the voltage regulator stops to regulate against the further reduction in input voltage is called the dropout voltage. This happens when the input voltage gets close to the output voltage. The voltage regulator operation can be explained by using the nMOS pass transistor I-V characteristics. There are two common operational regions for a CMOS transistor operating in the above threshold region: the linear region and the saturation region. In the linear region, the pass transistor acts like a tunable series resistor. In the saturation region, the pass transistor becomes a voltage-controlled current source. Voltage regulators are designed to work in the saturation region. When the series pass transistor is in the saturation region, it acts as a current source as a function of V_{gs}. Under different load conditions, V_{gs} of the pass transistor controls the regulator to supply the demand output load current. Fig. 5.3 shows the input-output characteristics of an LDO regulator. In this figure, the LDO regulaor begins dropping at $3.65V$ input voltage. The range of the dropout region is between $2V$ and $3.65V$. Below this, the circuit is not functional. Low dropout voltage is necessary to maximize the regulator efficiency.

B. Quiescent Current or Ground Current

The difference between the input and output currents is called the quiescent current, or ground current. In order to have maximum efficiency, the quiescent current must be minimized. Quiescent current is defined by

$$I_q = I_I - I_O \tag{5.1}$$

Quiescent current includes all bias currents (e.g. band-gap reference, the resistor divider, and error amplifier) and drive current of the series pass transistor, which does not contribute to the output power. In CMOS voltage regulators, this drive current is zero.

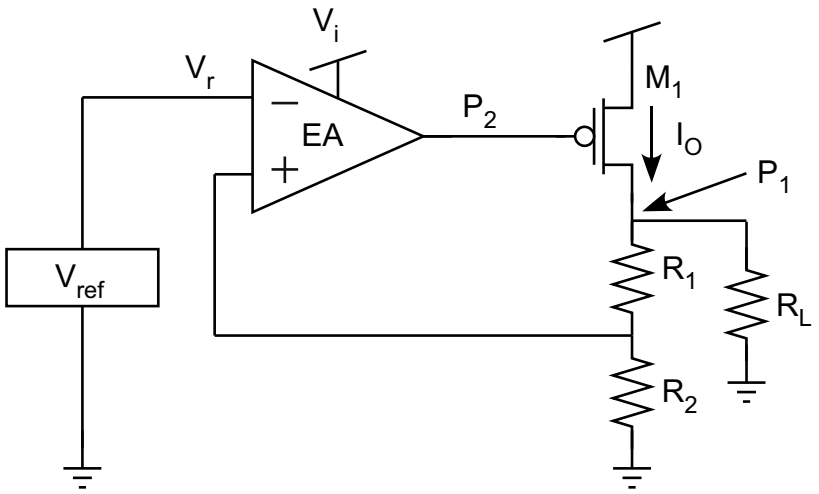


Figure 5.4: A PMOS Voltage Regulator

C. Efficiency

The efficiency of an LDO regulator is limited by the quiescent current and input/output voltages, and it can be expressed as:

$$\eta = \frac{I_o v_o}{(I_O + I_q) v_I} \times 100\% \tag{5.2}$$

In order to increase the efficiency of the LDO, the drop-out voltage (V_{do}) and the quiescent current must be minimized.

D. Load Regulation

Load regulation is the circuit's ability to keep the output voltage at a specific value under different load currents. The following expression defines the load regulation of a regulator:

$$LoadRegulation = \frac{\Delta V_o}{\Delta I_o} \tag{5.3}$$

Fig. 5.4 shows a PMOS voltage regulator. The output voltage change (ΔV_o) for a given load current change (ΔI_O) can be calculated as follows. Assume that M_1 is the series pass transistor, and g_m is the current gain of M_1, A is the voltage gain of the amplifier at its operating point. Assume that there is a small output current change (ΔI_o). The output voltage change can be expressed as

$$\Delta V_o = \Delta I_o R_{eq} \tag{5.4}$$

where R_{eq} is the equivalent output resistance at the output terminal:

$$R_{eq} = (R_1 + R_2)||R_L \approx R_L \tag{5.5}$$

The change of sensed voltage multiplied by the gain of the feedback path must be high enough to

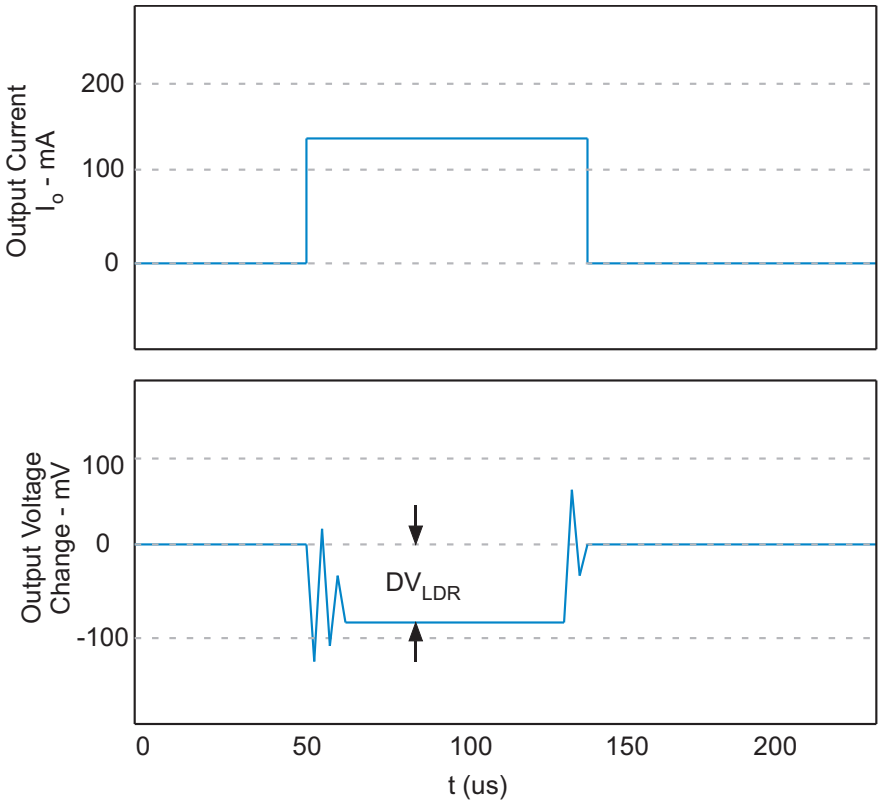


Figure 5.5: Load Transient Response of a Voltage Regulator.

achieve the specified change at the output current. Thus,

$$\Delta I_O = g_m A \frac{R_2}{R_2 + R_1} \Delta V_O \tag{5.6}$$

The load regulation can be obtained from the following expression:

$$\frac{\Delta V_O}{\Delta I_O} = \frac{1}{g_m A} \left(\frac{R_1 + R_2}{R_2} \right) \tag{5.7}$$

Therefore, increasing dc open-loop current gain improves load regulation. The worst case of the output voltage variations occurs as the load current changes from zero to maximum allowable I_d or vice versa. Fig. 5.5 shows a load transient response of a voltage regulator. The load current starts at zero and it suddenly jumps up to $100mA$. The voltage drop (ΔV_{LDR}) visible in the output determines the load regulation of the regulator.

E. Line Regulation

Line regulation is a measure of the circuit's ability to keep the specified output voltage with varying input voltage and is expressed as:

$$LineRegulation = \frac{\Delta V_O}{\Delta V_I} \tag{5.8}$$

The output voltage variations for a given input voltage change can be calculated from Fig. 5.4 as

follows

$$V_O = \frac{V_I R_{eq}}{R_{ds} + R_{eq}} - \Delta V_O = \frac{V_I R_{eq}}{R_{ds} + R_{eq}} - \Delta I_O R_{eq} = \frac{V_I R_{eq}}{R_{ds}} - G(V_s - V_r)R_{eq} \tag{5.9}$$

Where the open loop current gain $G = \beta \times g_a$, and R_{ds} is the equivalent resistor between drain and source of the series pass element, and R_{eq} is the equivalent output resistance at the output terminal:

$$R_{eq} = (R_1 + R_2)||R_L \approx R_L \tag{5.10}$$

And, the sensed output voltage is given by

$$V_s = \frac{R_2}{R_2 + R_1} V_O \tag{5.11}$$

Now, substituting (5.11) in (5.9), and assuming that $GV_s \gg 1$, we have:

$$V_O = \frac{R_1 + R_2}{GR_2(R_{ds} + R_{eq})} V_I + \frac{R_1 + R_2}{R_2} V_r \tag{5.12}$$

This equation has two parts, the first one is the steady-state average output voltage, and the other one is a function of the input voltage. Therefore, the line regulation can be expressed as:

$$\frac{\Delta V_O}{\Delta V_I} = \frac{1}{(R_{ds} + R_L \beta g_a)} \frac{R_1 + R_2}{R_2} \tag{5.13}$$

This equation shows that increasing the dc loop current gain improves the line regulation.

F. Transient Response

The transient response is the maximum allowable output voltage variation for a step change in the load current. The transient response is a function of the output capacitor value (C_O), the equivalent series resistance (ESR) of the capacitor, the bypass capacitor (C_b) (Fig. 5.6), and the maximum load-current (I_{O-Max}). The maximum transient voltage variation is defined as [71]:

$$\Delta V_{tr,max} = \frac{I_{O,max}}{C_O + C_b} \Delta t_1 + \Delta V_{ESR} \tag{5.14}$$

where Δt_1 corresponds to the closed loop bandwidth. ΔV_{ESR} is the voltage variation resulting from the presence of the ESR (R_{ESR}) of the output capacitor. ΔV_{ESR} is proportional to R_{ESR}.

G. Frequency Response

Fig. 5.7 shows a small-signal model of the linear regulator. The transconductance is modeled by g_a with a load comprised of capacitor C_{par} and R_{par}. The series pass element is modeled by a small-signal model with transconductance g_p. An output capacitor (C_O) is shown with an equivalent

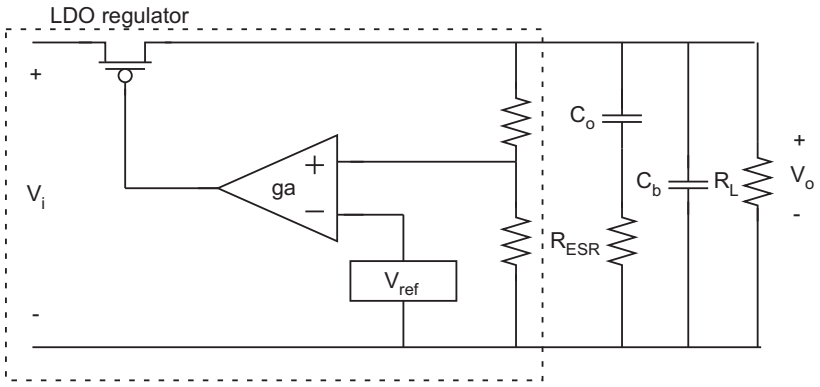


Figure 5.6: A PMOS Voltage Regulator

series resistor (R_{ESR}). The bypass capacitor is C_b. The output impedance of this structure is:

$$Z_O = R_{12p}||(R_{ESR} + \frac{1}{SC_O})||\frac{1}{SC_b} = \frac{R_{12p}(1 + SR_{ESR}C_O)}{S^2R_{12p}R_{ESR}C_OC_b + S[(R_{12p} + R_{ESR})C_O + R_{12p}C_b] + 1} \tag{5.15}$$

where $R_{12p} = R_{ds}||(R_1 + R_2) \approx R_{ds}$. The output load capacitor is usually much bigger than the bypass capacitor. Thus, the output impedance Z_O can be approximated as

$$Z_O \approx \frac{R_{ds}(1 + SR_{ESR}C_O)}{[1 + S(R_{ds} + R_{ESR})C_O] \times [1 + S(R_{ds}||R_{ESR})C_b]} \tag{5.16}$$

From this equation, some poles and zeros of the regulator are obtained. The first pole is obtained from this equation:

$$P_o = \frac{1}{2\pi(R_{ds} + R_{ESR})C_O} \approx \frac{1}{2\pi R_{ds}C_O} \tag{5.17}$$

The second pole is obtained from 5.16 again:

$$P_b = \frac{1}{2\pi(R_{ds}||R_{ESR})C_b} \approx \frac{1}{2\pi R_{ESR}C_b} \tag{5.18}$$

and the zero is:

$$Z_{ESR} = \frac{1}{2\pi R_{ESR}C_O} \tag{5.19}$$

In addition, another pole exists from the output impedance of the amplifier:

$$P_a \approx \frac{1}{2\pi R_{par}C_{par}} \tag{5.20}$$

Fig. 5.8 shows a possible frequency response of the LDO voltage regulator.

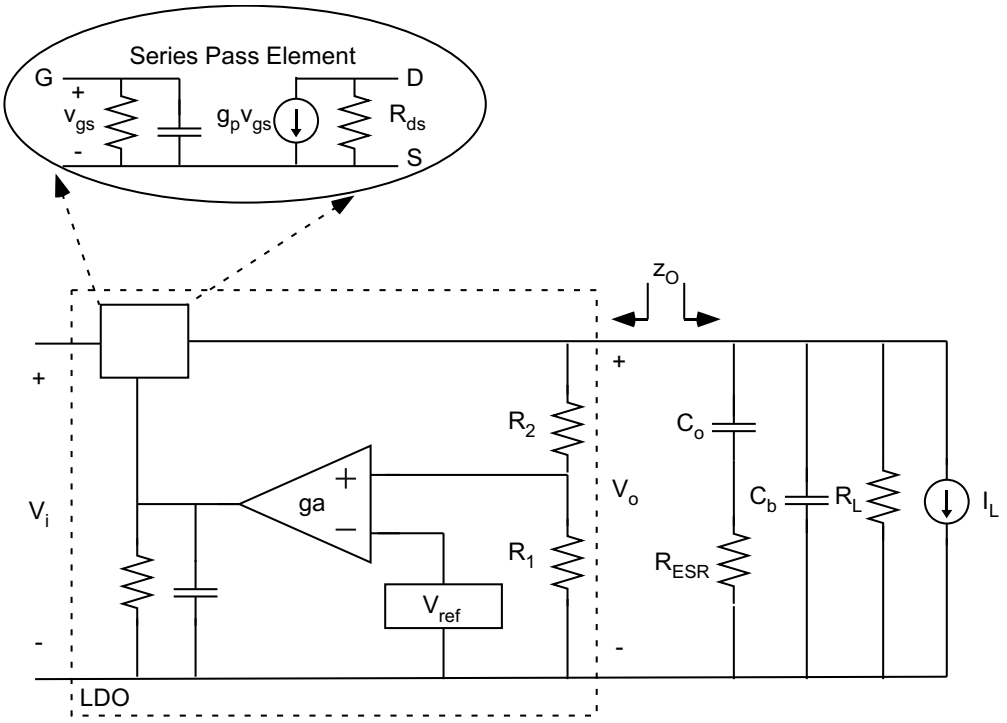


Figure 5.7: AC Model of a Linear Regulator

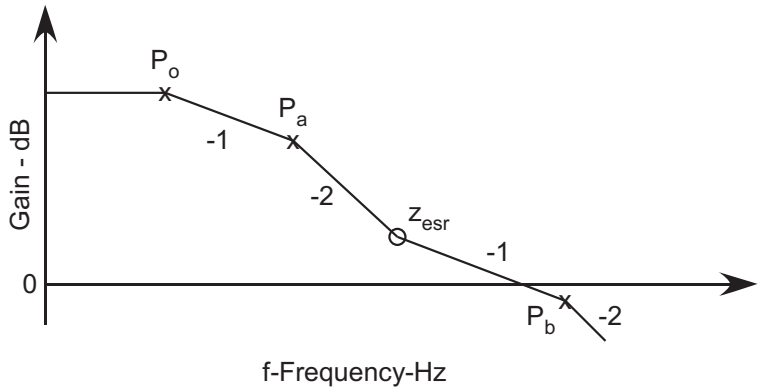


Figure 5.8: Frequency Response of the LDO Voltage Regulator

5.2 Analog/Digital LDO regulator

A portable system is usually powered by a battery. However, the supply voltage that is provided through the battery is noisy and will drop over long-term use. In this work, we use a 3.3V battery (input to the LDO regulator), and the reference voltage to the regulator is a 0.9V reference voltage. We use the voltage regulator presented in [72]. This LDO regulator includes a current-boosting circuit with fast on/off features. Therefore, it can momentarily provide an extra current to charge and discharge the parasitic capacitor at the gate of M_P. Thus, the voltage buffer provides a fast slew rate at the gate of the power transistor while the quiescent current of the LDO remains always low. The circuit diagram of the voltage regulator is shown in Fig. 5.9.

The operation of this circuit is explained in the remainder of this section. When I_O is suddenly

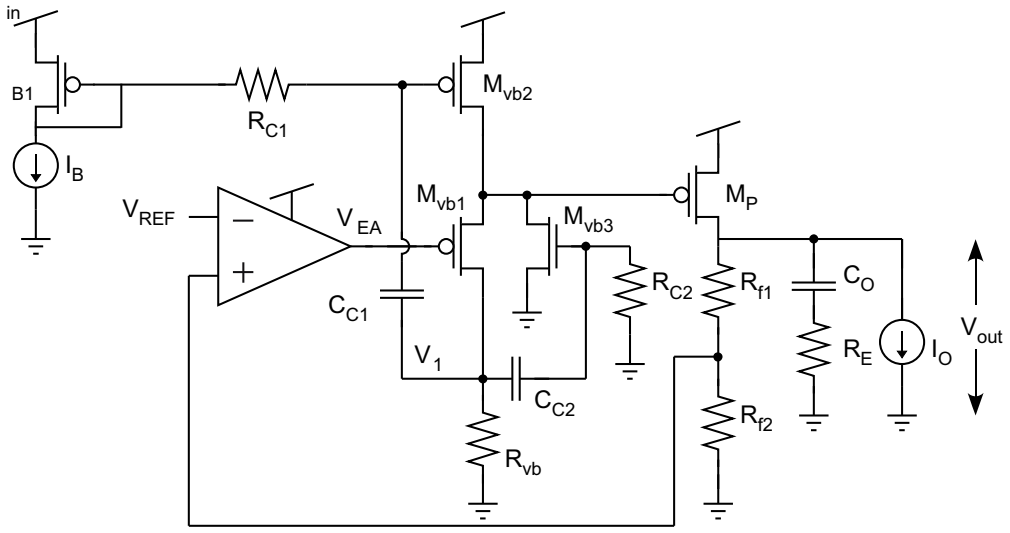


Figure 5.9: Schematic of the LDO regulator

reduced, an overshoot at V_O is generated. The amplifier detects this overshoot through the resistor divider and generates a voltage (V_{EA}). When V_{EA} increases, V_1 is decreased with a very large amplitude due to the high-gain common-source amplifier. C_{c1} couples this voltage change to decrease the gate voltage of M_{vb2}. Thus, M_{vb2} will conduct more current to charge the parasitic capacitor at the gate of M_P. During this time, M_{vb3} remains OFF and it doesn't affect the regulation.

When I_O is suddenly increased, it causes an undershoot in V_O. V_{EA} will decrease and V_1 increases. C_{C2} couples the voltage change of V_1 to the gate of M_{vb3}. Thus, this transistor turns on and provide an extra current to discharge the parasitic capacitor at the gate of M_P. The change of V_1 is also coupled through the capacitor, C_{C2}, to the gate of M_{vb2}, which turns off M_{vb2} momentarily because the gate-source voltage of this transistor starts to reduce.

5.2.1 Experimental results

This LDO regulator was designed and fabricated in a 0.35μm standard CMOS process. The supply voltage of this LDO regulator is at $3.3V$. The bias current (I_B) and the bias current for the amplifier are provided from an off-chip source. The load regulation plot of this regulator is shown in Fig. 5.10. A 10.5μF tantalum capacitor was used for this measurement. According to this plot, the voltage change from 0 to $10mA$ of the load current is $25mV$.

5.3 Mid-rail Regulator

A mid-rail regulator is a regulator that is used to make a clean AC ground or a mid-rail voltage [73, 74]. This voltage is half the supply voltage, which is 1.25V in our system. This voltage is usually used as a reference voltage to be compared with analog signals and create digital signals. Unlike a typical voltage regulator, this regulator must support both sink and source load currents.

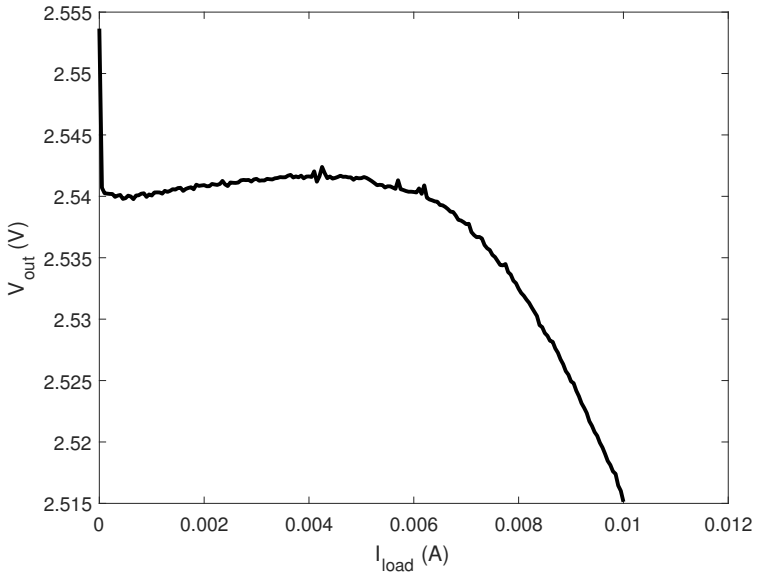


Figure 5.10: Load regulation plot of the voltage regulator

The load current is in the range of few microamps. Fig. 5.11 shows the circuit diagram of this regulator. This regulator includes two identical OTAs shown in Fig. 5.12. The supply voltage of this circuit is a 3.3V battery. This OTA is biased in the subthreshold region. The bias current (I_{bias}) of this OTA is 120nA. The transistor sizes of this OTA are as follows:

$$(W/L)_{Mb1} = (W/L)_{Mb2} = \frac{20\mu m}{5\mu m} \tag{5.21}$$

$$(W/L)_{M1} = (W/L)_{M2} = \frac{10\mu m}{5\mu m} \tag{5.22}$$

$$(W/L)_{M3} = (W/L)_{M4} = (W/L)_{M5} = (W/L)_{M2} = (W/L)_{M6} = (W/L)_{M7} = \frac{10\mu m}{5\mu m} \tag{5.23}$$

Therefore, each OTA consumes 360nA in total and the regulator consumes 720nA in total. To make two separate control loops, the positive (negative) input of the OP_1 and OP_2 are tied together. Also, both OP_1 and OP_2 see the gate of the pass transistors (M_1 and M_2) as their load. Therefore, V_x and V_y have almost the same voltage. The source terminals of M_1 and M_2 are tied to V_{out}. Thus, only one transistor is on and the other transistor is in the cut-off region. This will prevent the short current which is the result of having both of these transistors "on" at the same time.

Now, assume that we have a sink load current and V_{out} is initially greater than 1.25V. Therefore, both OP_1 and OP_2 are saturated to the low level. Thus, M_2 is on and M_1 is in the cut-off region. In this state, the upper loop is open and the bottom loop adds some more current to the sink load current. This will discharge the parasitic capacitors connected to the V_{out}. The discharging process continues until V_{out} gets below 1.25V. At this point, $V_x = V_y$ switches to a higher voltage because the positive input of the amplifier is higher than the negative input of the amplifier. At this point

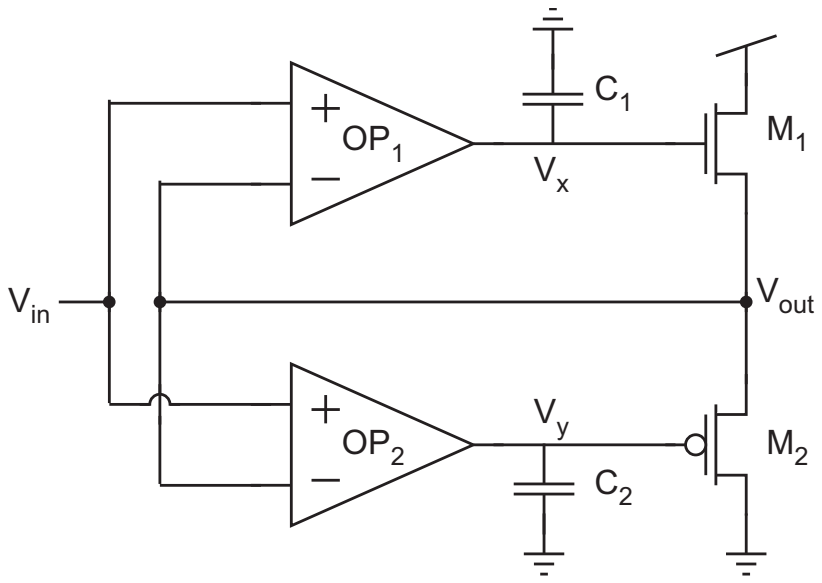


Figure 5.11: Circuit diagram of the mid-rail regulator

OP_1 and M_1 make a regulation loop and will make V_{out} close to 1.25V. If V_{out} is initially less than 1.25V, then M_2 will be in the cut-off and M_1 charges V_{out} and makes it closer to 1.25V. The same analysis can be applied to a situation with a source load current.

Fig. 5.13 (top) shows the measured line regulation of this regulator when the supply voltage sweeps from 2.5V to 3.3V. Fig. 5.13 (bottom) shows the measured current that is consumed by the circuit. The difference between the predicted value and measured result is around 30nA, which is due to the accuracy of the bias current and mismatch of the current mirrors. The maximum voltage change in the output is around 1.5mV. Therefore, the line regulation can be expressed as:

$$LineRegulation = \frac{\Delta V_O}{\Delta V_I} = \frac{1.5mV}{0.8V} = 1.8mV/V \tag{5.24}$$

The same measurement was done for $-10\mu A$ and $10\mu A$ of load current. Fig. 5.14 and 5.15 show the measurement results of this regulator for $-10\mu A$ and $10\mu A$ of load current. The LR is around 1.6mV/V and 3.1mV/V for $10\mu A$ and $-10\mu A$ of load current, respectively.

Fig. 5.16 shows a load sweep measurement on the mid-rail regulator. The load current range of this regulator is $-10\mu A \sim 10\mu A$. The output voltage difference between positive $I_{load} > 0\mu A$ and negative load currents $I_{load} < 0\mu A$ is due to the difference that exists between the loop gain of the top and bottom loops. At first, it seems that these two voltages must be the same because we are using the same OTAs with matched layouts. However, the difference between the gains of common-drain structures created with M_1 and M_2 generates a voltage difference between these two cases.

Finally, the PSRR of the mid-rail regulator is measured and shown in Fig. 5.17 and 5.18. For this measurement, a dynamic signal analyzer (DSA) was used. A $100mV$ Sine wave was applied

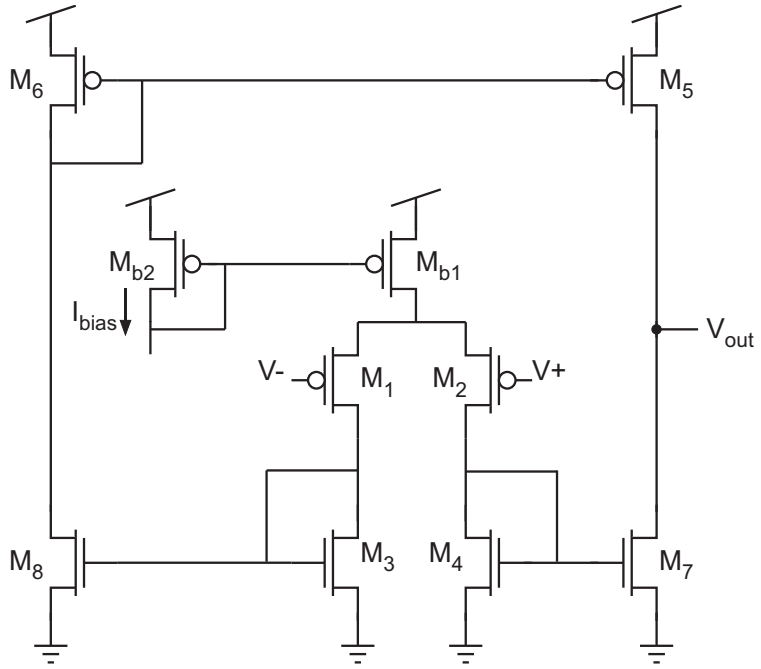


Figure 5.12: Circuit Schematic of the OTAs used in the mid-rail regulator.

to the supply voltage and the frequency of the signal was swept from DC to 100KHz. Finally, the amplitude of the V_{out} was measured through the DSA. To measure the PSRR, the following equation is used:

$$PSRR = dB(100mV) - dB(V_{out}) \tag{5.25}$$

These figures show that the minimum PSRR is around 30dB.

5.4 Conclusion

In this chapter, two types of voltage regulators are presented using a 0.35μm CMOS process. The first voltage regulator is designed for analog/digital power supplies. This circuit is supposed to receive a DC voltage at the input and generate a low-noise DC signal for analog/digital supplies. The experimental results of this voltage regulator are also presented. The second voltage regulator is presented for the mid-rail voltage ($1.25V$) of our system. This voltage regulator must provide both source and sink load currents.

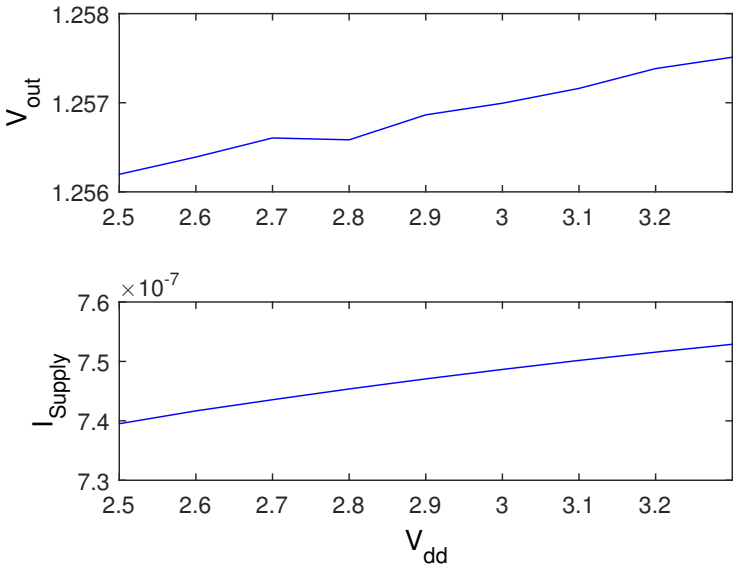


Figure 5.13: (a) Line Regulation of the mid-rail voltage regulator for 0 load current (b) Current Consumed by mid-rail voltage regulator for 0 load current

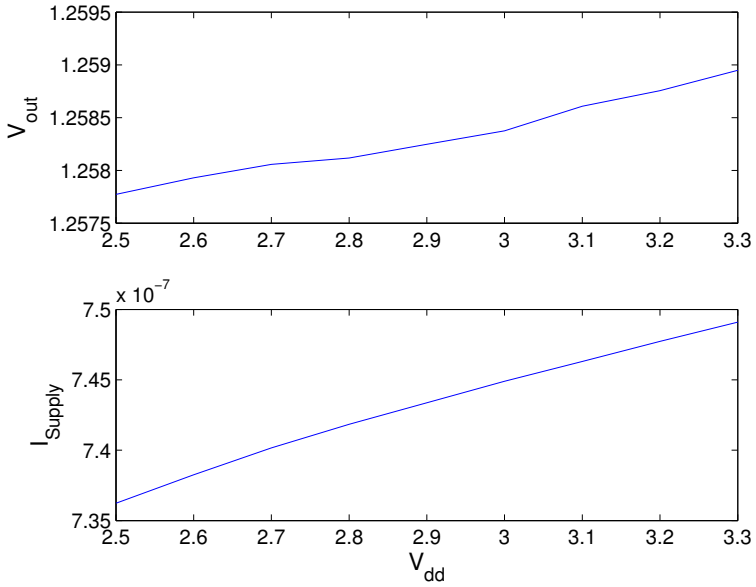


Figure 5.14: (a) Line Regulation of the mid-rail voltage regulator for $10\mu A$ load current (b) Current Consumed by mid-rail voltage regulator for $10\mu A$ load current

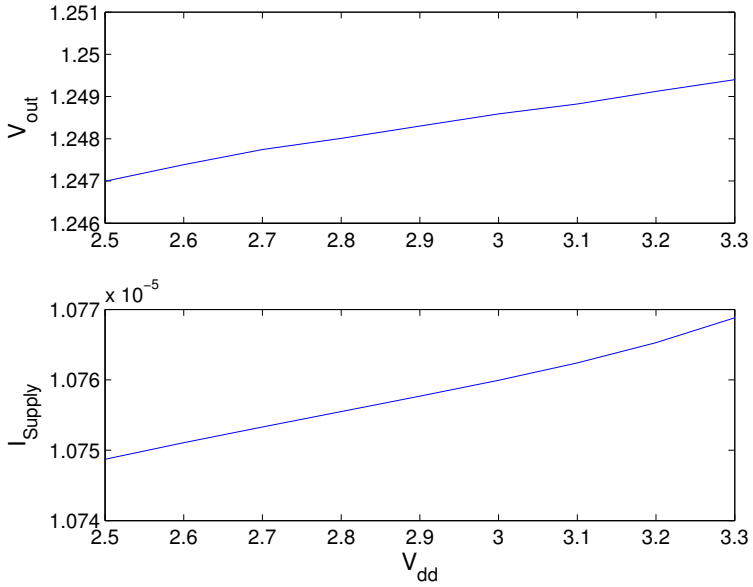


Figure 5.15: (a) Line Regulation of the mid-rail voltage regulator for $-10\mu A$ load current (b) Current Consumed by mid-rail voltage regulator for $-10\mu A$ load current

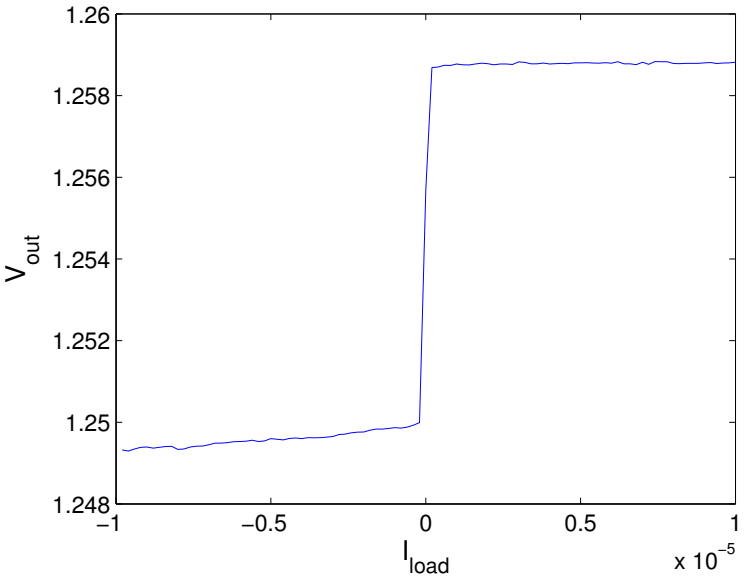


Figure 5.16: Load Sweep of the mid-rail voltage regulator

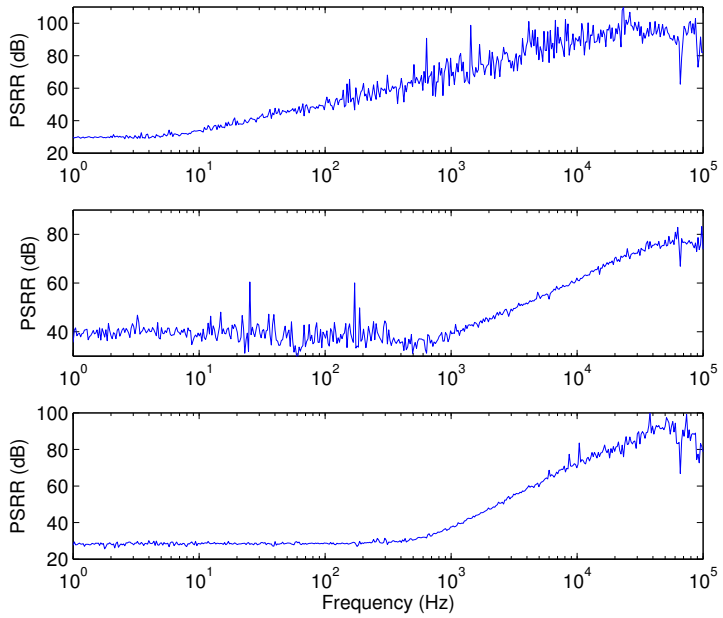


Figure 5.17: PSRR of the mid-rail regulator under 2.5V supply for (top) 0 load current (middle) $10\mu A$ load current (bottom) $-10\mu A$ load current

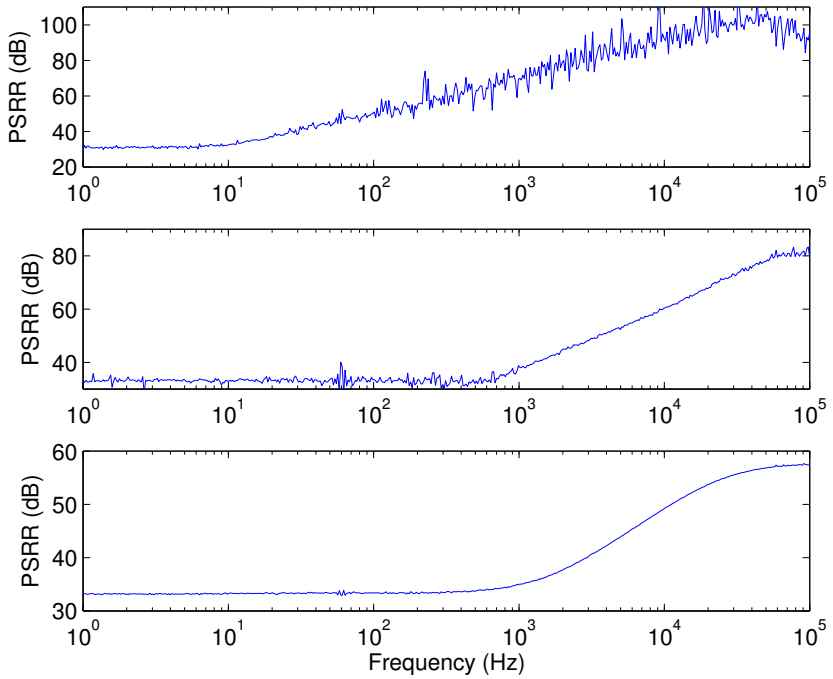


Figure 5.18: PSRR of the mid-rail regulator under 3.3V supply for (top) 0 load current (middle) $10\mu A$ of load current (bottom) $-10\mu A$ of load current

Chapter 6

Below-Ground Injection of FG Transistors for Programmable Analog Circuits

FG transistors have been used in many systems to add non-volatile memory to the system. By storing a precise amount of charge on the electrically isolated floating gate, these devices can be used in analog applications as programmable current sources. However, as we have seen in previous chapters, the process of selecting an individual FG and precisely programming it requires significant infrastructural overhead, especially when considering that both tunneling and injection require the on-demand generation of voltage drops greater than V_{dd}. In this chapter, a novel technique for simplifying the programming of floating gate transistors is presented. Specifically, we present a technique to use "below-ground injection" that permits injection of individual FG transistors in a large array but does not require any isolation switches to operate below the substrate voltage. The need to have selection circuitry operating at negative voltages is circumvented by using indirectly programmed FG transistors and a circuit that linearizes the programming currents. Consequently, this technique can be used in any standard single-well CMOS process to accurately program analog values on FG transistors. Also, two charge pumps that can produce the negative voltages are presented here. The end result permits an overall system that requires far less infrastructure and power consumption than a more conventional system using voltages above V_{dd} to create injection conditions. All results were measured from an integrated circuit fabricated in a standard 0.5μm CMOS process.

6.1 Below Ground Programming

Hot-electron injection can be used to add electrons to the floating gate. The voltages required for injection typically exceed the rated V_{dd} of the process but are generally lower than junction

breakdown voltages. Consequently, standard transistors can be used to isolate/multiplex a specific FG for injection. Because of the ability to select a specific FG for injection, injection has become the preferred method for programming large arrays of FGs in analog applications [53]. PMOS devices have dominated analog floating gate circuits because the injection currents necessary for programming can be created much more efficiently in PMOS devices than NMOS devices. The amount of injection current in a PMOS device can be represented by the simplified equation:

$$I_{inj} \approx \beta {I_s}^{\alpha} e^{V_{sd}/V_{inj}} \tag{6.1}$$

where α, β, and V_{inj} are fit parameters. This equation covers the subthreshold operating range, which is where the injection operation is most efficient. Of note, the channel current (I_s) and the source-to-drain voltage (V_{sd}) determine the injection rate. As was mentioned earlier, significant injection only occurs for $V_{sd} > V_{dd}$.

When injecting a floating-gate transistor to a specific analog value, the channel current and V_{sd} must be precisely controlled. The technique of simply raising the source voltage to initiate injection is not ideal; this would alter V_{sg}, thereby changing the channel current to an unknown value. This simple technique makes it hard to achieve precise programming results.

Instead, the more conventional approach is to lower the drain potential. Because V_{sg} stays constant and because the drain has little influence on the channel current (due to its output resistance), the channel current remains at a known value. Lowering the drain potential below ground seems to be an obvious choice to generate the large V_{sd} required for injection; however, this has long been difficult to implement in the most conventional system. For example, when working with arrays of FG transistors, a multiplexer at the drain of the FG transistor is needed to connect the drain to a negative voltage in "program mode" but then connect the drain to its circuit during a "run mode" [2]-[53]. In a single-well process, making selection circuitry operate below ground is not easily achieved, since the NMOS devices in a standard transmission gate would have their diffusion areas forward biased.

Instead of using negative voltages for injection, the common procedure for injecting FGs for analog applications is to raise the source and gate voltage up well above V_{dd} in a fashion that permits the channel current to be known. Then, the drain can be at a positive voltage and still provide a V_{sd} that is large enough to invoke injection. This process of "ramping up" all voltages associated with the floating gate transistors and then lowering the drain to start injection, requires significant infrastructure. Particularly if all of the programming infrastructures is to be included on-chip, the infrastructure should be made to be compact, low-power, and easy-to-use.

In the next section, we present a simple circuit that permits below-ground injection, thereby greatly simplifying the necessary infrastructure. This circuit also eliminates the need for multiplexers at the drain of the FG transistor undergoing injection. In section 6.1.2, we discuss a

charge-pump circuit that can be used to generate the negative voltages on-chip, and in section 6.1.3, we show the results of this circuit and describe how it can be used in a large array.

6.1.1 Floating-Gate Memory Cell

In [2], we described an FG memory cell that can be used to linearize the injection process, and a simplified version of this technique is illustrated in Fig. 6.1(a). An amplifier is used to modulate the control gate voltage, V_{cg}, of the FG transistor so that the floating node maintains at a constant voltage during injection and/or tunneling—this negative feedback ensures that the source-to-floating-gate voltage remains constant so that I_{ref} continually flows through the device. In [2], I used "above-ground" injection to program the FG transistor, by raising V_{dd} and then lowering the drain of the FG transistor toward the ground. However, this technique still suffers from needing to first "ramp up" the FG transistor before programming, and the associated infrastructure to do so.

Fig. 6.1(b) illustrates a simplified version of the memory cell proposed in this work. This new memory cell is an extension of [2] with the addition of a second transistor connected to the FG node. The current flowing through $M_{circuit}$ is directly set by the stored charge on the FG, and it is connected directly to a circuit to provide a bias current. All injection current goes through M_{inj}, so $M_{circuit}$ never needs to experience large V_{sd} values. Instead, the drain of M_{inj} is lowered below ground to induce injection. A multiplexer at the source of M_{inj} is used to disconnect M_{inj} from the feedback amplifier when the transistor is not actively being programmed. As a result, many FG transistors can all be connected to a single drain node that is lowered below ground, and injection can be prevented in a non-selected FG by connecting its source to ground. This technique permits the selection of an individual FG transistor for programming, and no selection switch needs to operate below the substrate potential.

Fig. 6.2 shows a block diagram of an FG array when a specific memory cell is being programmed. In this method, the drain connection of all M_{inj} transistors in all memory cells will be connected to the output of a negative charge pump. The charge pump is designed such that the output voltage is either at V_{dd} when disabled or a negative value when enabled. When a specific memory cell requires injection, the multiplexer inside the target memory cell will be configured such that the memory loop is closed. At the same time, the source connection of all other M_{inj} transistors in the rest of the array will have their source connected to ground by the correctly set-up of the multiplexers inside the memory cells. Therefore, when a negative voltage is applied to $V_{negative}$ only one transistor has a V_{sd} large enough to generate injection.

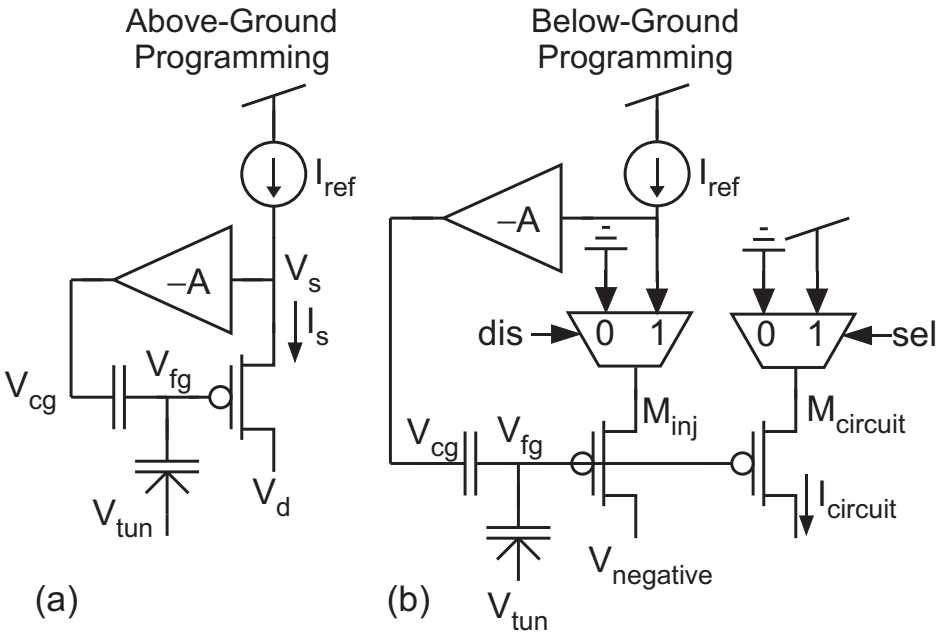


Figure 6.1: (a) Continuous programming method we presented in [2] using voltages above V_{dd}. (b) The proposed Continuous programming method employing negative voltages.

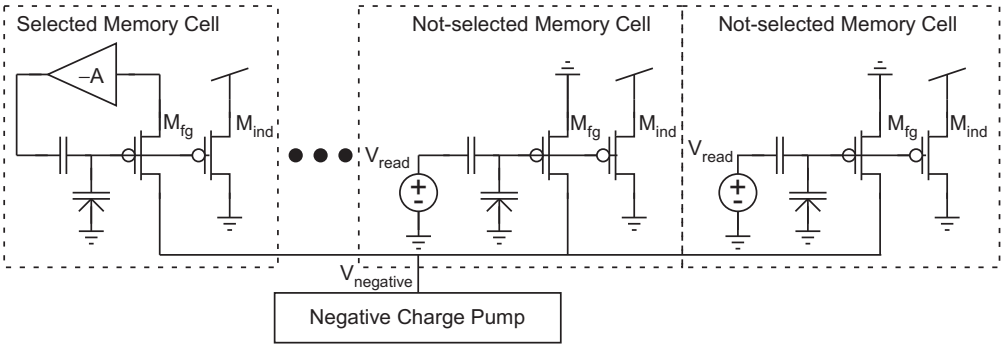


Figure 6.2: Proposed technique to program a selected transistor in an array of FG transistors using a negative charge-pump

6.1.2 Negative Charge Pump

As I mentioned in Chapter 2, hot-electron injection is a conventional method to program the conventional flash memory cells (e.g. NOR Flash). In this approach, a high voltage ($\sim 6.5V$) is applied to drain, while a high voltage is applied to the gate ($\sim 10.5V$) and the source terminal is grounded. For the erasure, a high positive voltage ($\sim 10.5V$) is applied to the source while its gate is grounded [75]. Applying a very high voltage to the source for erasure may cause reliability issues because this high voltage will reverse-bias the source-to-bulk PN junction close to the breakdown region. Also, the band-to-band tunneling and the high electric field between source and substrate can create hot holes and trap them in the gate oxide leading to degrading cell endurance [75]. A more reliable approach to create the electric field across the gate oxide is to apply a high negative voltage to the gate (-8V to -12V), while keeping the source typically at V_{DD}. For many

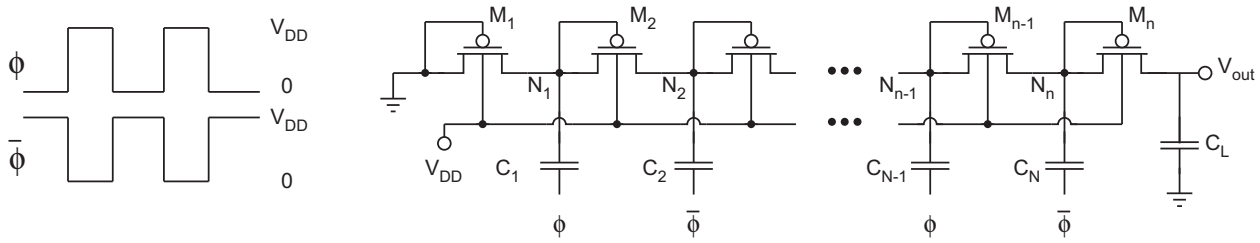


Figure 6.3: A Dickson charge pump used for negative voltage generation

portable applications such as RFID tags, mobile phones, and digital cameras which use non-volatile memories, high voltages are necessary for program and erase operations. For these applications, it is useful to be able to generate all these voltages inside the chip from the power supply of the portable system [76, 4]. The outputs of these circuits must be handled with the read voltage to feed the memory Word Line (WL) with the required level under any operating conditions (reading, erasing, programming) [75].

Traditionally, two types of charge pumps (negative and positive) are realized on a chip. However, only one was active at a time due to the timing of the program and erase operations [77, 78, 75, 5]. In some other designs, a single charge pump is used as a switched polarity charge pump. [79] presents an example of a switched polarity charge pump.

In Chapter 3, three positive charge pumps were presented for tunneling and injection of the FG transistors. In this section, a negative charge pump is presented which is suitable for the injection process. A review of the available negative charge pumps is presented in the remainder of this section. Then, the proposed negative charge pump is discussed.

Many charge pump circuits are based on a well-known structure called the Dickson charge pump [33, 80]. In these charge pumps, MOS transistors are connected like a diode, so the charge can travel in just one direction. Fig. 6.3 shows a Dickson charge pump to generate negative voltages. Unlike positive charge pumps which use V_{DD} as the input signal, the input of the negative charge pump is connected to ground. C_1 to C_N are coupling capacitors with the same value. Two opposite phase pumping clocks with the same amplitude (V_{DD}) are used to control the operation. When a transistor turns on, it will enter saturation region. In Fig. 6.3, the first transistor is always on, which makes N_1 to be at ground. When ϕ is low and $\overline{\phi}$ is high, the second transfer transistor is turned on which makes N_2 be 0V. When ϕ is high and $\overline{\phi}$ is low, the second transistor turns off and the bottom plate of C_2 goes to ground. This will shift N_2 to $-V_{DD}$. The same procedure makes the output of the next stage to be $-2V_{DD}$. Ideally, the output of a N stage negative Dickson charge pump is $-NV_{DD}$ in steady state. In this analysis, we ignored the threshold voltage of the transistors. However, in reality, when a diode-connected transistor turns on, it goes into saturation region and the voltage drop across the transistor is V_{th}. The threshold voltage of the PMOS transistors across the diode-connected transistor chain increases from the input to the output due to the body effect. This is because the body of PMOS transistors are connected to V_{DD} and the body-source voltage

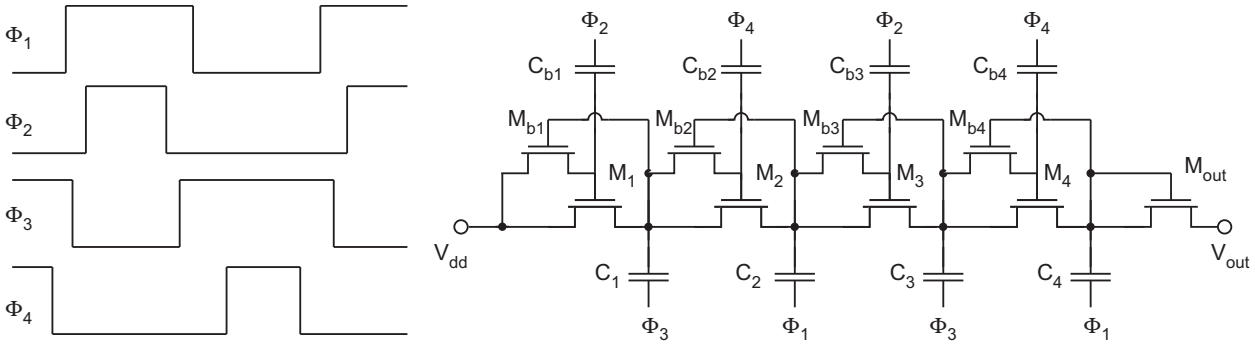


Figure 6.4: Circuit structure of the boosting charge pump [4]

of the transistors keep increasing as it gets closer to the final stages. The voltage gain at each stage is reduced by this voltage drop across the switch transfer transistors.

[4] uses a charge transfer gate with threshold voltage cancellation. This structure fixes the threshold voltage loss of the transfer gates and increases the efficiency at the expense of a more complex clocking structure. This structure was used for positive output voltages, but it is applicable to negative charge pumps too. The circuit diagram of this structure is shown in Fig. 6.4. In this new architecture, boosting capacitors ($C_{b1} \sim C_{b4}$) and boosting transistors ($M_{b1} \sim M_{b4}$) are added. Using this method, the gate potential of the charge transfer transistors ($M_1 \sim M_4$) can raise to a higher level than the drain terminal. This makes the transistor fully open by biasing the transistor in the linear region. Therefore, the drain and source voltages can reach the same level. This approach is called threshold voltage cancellation scheme [79]. The negative four-phase charge pump improves the voltage gain of the charge pump; however, because of the fixed body potential (e.g. ground in [79]) of the transistors, this structure still suffers from threshold rising problem. Both the substrate leakage current issue and the body-effect can be handled in positive charge pumps by using either both P-type and N-type CTSs in a triple well process, or only using P-type CTSs in an N-well process [5]. The issue of implementing a negative charge pump in N-well and triple well processes are shown in Fig. 6.5 (a) and (b), respectively. According to this figure, a negative charge pump can only get rid of both body-effect and substrate leakage current problems by adopting N-type CTSs in a triple-well process. For example, [81] is a negative charge pump generating -9.5V for the wordline of a NOR flash memory which uses intrinsic n-channel transistors in a triple well. In [81] during first pumping cycle, ϕ_2 brings N_1 to low level. At the same time, ϕ_1 is high, which makes T_{b1} and T_2 turn off. Since N_2 is coupled high, transistor T_{b2} turns on and node M_2 is precharged. Now, during the boost cycle, the boost clock (ϕ_{b2}) goes to a high level. This will couple the precharge node (M_2) to a higher voltage level, and T_1 passes the negative charge from node N_1 to node N_2.

In the next pumping cycle, ϕ_1 couples node N_2 to a lower level. This makes the precharge transistor (T_{b2}) and transistor (T_1) off, while transistor T_{b1} is turned on and precharges M_1. The boost clock (ϕ_{b2}) increases the precharged node (M_2) to a higher voltage. Therefore, transistor T_2 passes the negative charge from N_2 to the next stage. Simultaneously to this cycle, ϕ_2 couples node

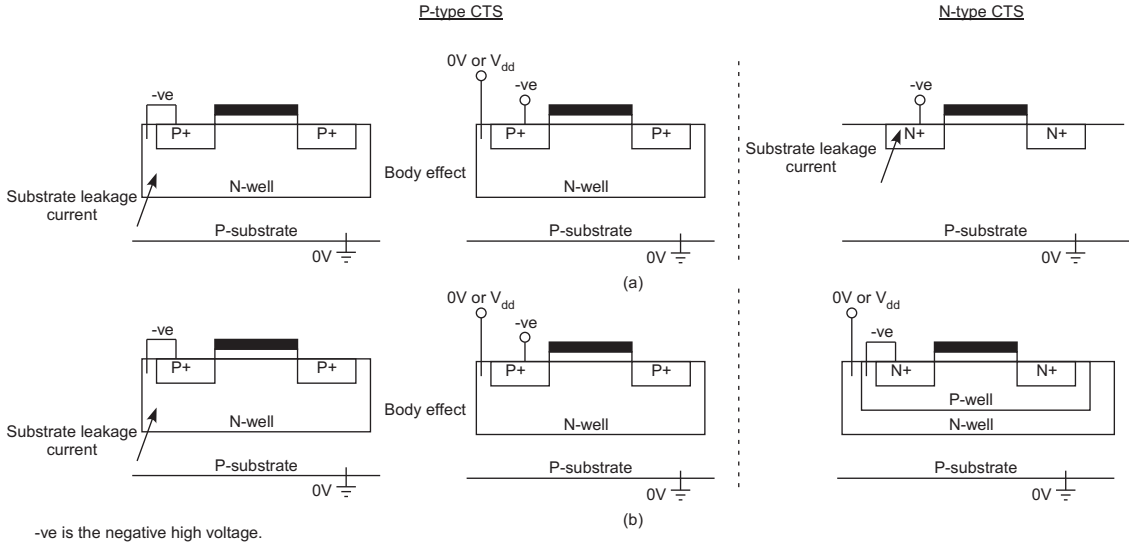


Figure 6.5: Implementation of a P-type and N-type transistor in a negative charge pump and their issues in (a) N-well process (b) Triple-well process [5].

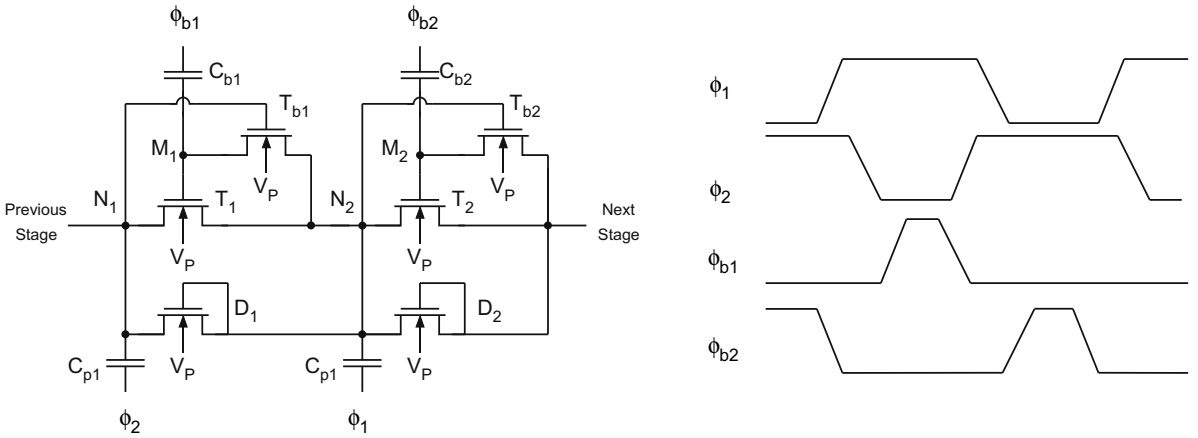


Figure 6.6: Circuit diagram of the negative charge pump presented in 6.5

N_1 to a higher level and causes a forward bias V_{BE} of the emitter base diode of the parasitic n^+pn bipolar transistor within the triple well [6].

Due to all of the issues mentioned above, some special circuits must be used to properly bias the P/N-well of the transistors in a negative charge pump. Fig. 6.7 shows the circuit diagram of the charge pump presented in [6]. The charge transfer transistors (T_1 and T_2) and boost transistors (T_{b1} and T_{b2}) function the same as Fig. 6.6, and the same clock scheme is used. The n-well of the charge pump is connected to 0V. The NMOS transistors T_{c1} and T_{w1} are used for keeping the p-well at the lowest voltage level within the charge pump stages. The drain of T_{w1} and T_{c1} are connected to $N2$ and $N1$, respectively. The source terminals of both of these transistors are connected to the common p-well of the pump stages. The gates of T_{c1} and T_{w1} are crosswise connected to $N2$ and $N1$, respectively. When $N1$ is coupled high, T_{w1} turns on. This happens when ϕ_1 and ϕ_2 are both high. When in charge pumping phase, T_{c1} does not contribute to the charge transfer to and from

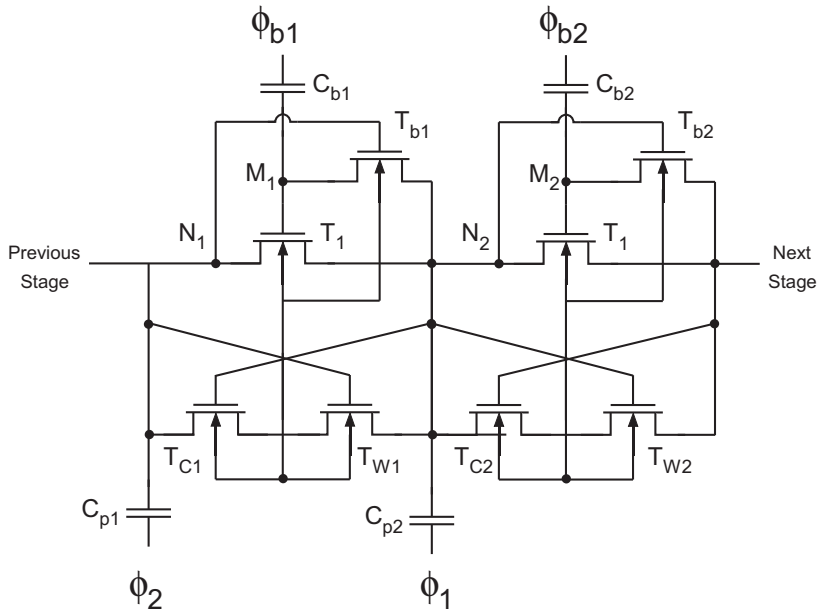


Figure 6.7: Circuit diagram of the negative charge pump presented in [6]

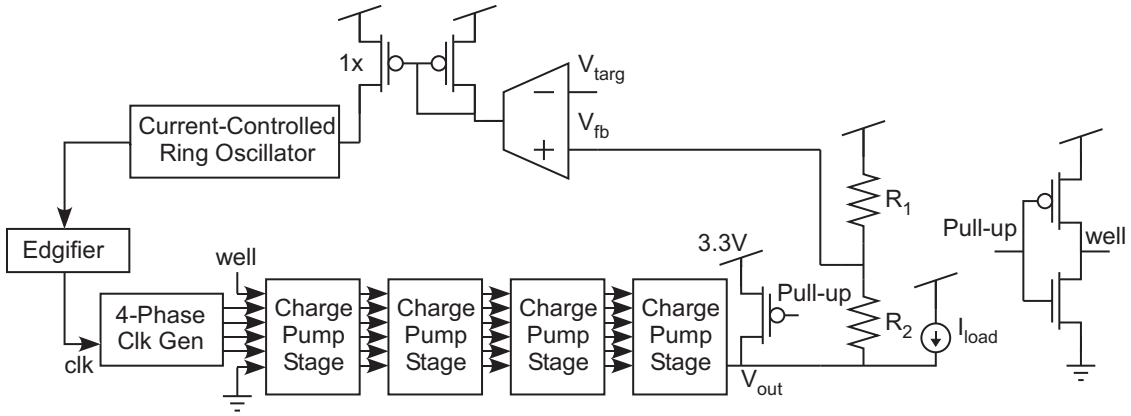


Figure 6.8: Block diagram of the negative charge pump.

the p-well. However, it reduces the p-well voltage below the voltage at the nodes $N1$ or $N2$ just by capacitive coupling.

There are some other negative charge pumps designed to avoid the increment of voltage drop between source and substrate in the PMOS charge transfer transistors [82, 78].

A generic block diagram of the regulated charge pump that generates the negative voltage for our programmer is shown in Fig. 6.8. This charge pump is based on [3] which uses a variable-frequency regulation technique to provide a low-ripple output voltage. Our charge pump uses 4 charge transfer switch stages as shown in Fig. 6.9. A $0.5\mu m$ n-well process is used for this circuit, and charge transfer switches are realized with PMOS transistors. In this charge pump, the enable input of the oscillator and Pull-up node are tied together. This will set the charge pump output to the chip supply voltage (3.3V) when the charge pump is disabled.

The n-well voltage (*well*) is controlled by the output of an inverter from which its input is

coming from an external pin ($Pull-up$). When charge pump is enabled, $Pull-up$ is set to the higher logic level (3.3V). This will make the n-well voltage ground through "*well*". When the charge pump is enabled, the negative voltages at source and drain terminals of the PMOS charge transfer switches make the PN junction reverse biased. When charge pump is enabled, $Pull-up$ is pulled low, which makes the charge pump output a positive voltage. "*well*" is pulled high, which makes the PN junction reverse biased again. To close the regulation loop in regulated negative charge pumps, [14] and [83] use a resistive divider to shift the negative voltage up to positive voltages. [84] uses a special type of comparator which compares the absolute value of the negative voltage with a positive voltage. In this work, we use the voltage divider approach to shift the output voltage up to positive values, and the difference between this voltage and target voltage (V_{targ}) is converted to a current. This current modulates the frequency of the current-controlled ring oscillator. The relation between V_{targ} and V_{out} can be derived by writing a KCL at the V_{fb} node:

$$V_{out} = \left(1 + \frac{R_2}{R_1}\right) V_{targ} - \frac{R_2}{R_1} V_{dd} \tag{6.2}$$

In this equation, we assume that the supply voltage is 3.3v and R_1 and R_2 are 300kΩ and 600kΩ, respectively. The schematic of the OTA is shown in Fig. 6.10. Ring Oscillator, Edgifier, and non-overlapping clock generator use the same circuit diagram as the circuits used in the charge pumps presented in Chapter 3. Fig. 6.11 shows the load regulation measurement results. This measurement was performed for output voltages from -5V to 1V and the current range of $100nA$ to 100μA. The charge pump stages are unable to provide enough negative charge at higher load currents, especially when output voltage become more negative. This is because the loop gain of the charge pump drops at higher load currents. Fig. 6.12 shows the transient response of the charge pump under $1\mu A$, $5\mu A$, $10\mu A$, $15\mu A$, and $25\mu A$ load currents. The charge pump is enabled first at $t = 0ms$. It takes around $0.5ms$ in the worst case (e.g. $25\mu A$) to reach the target voltage (-2.5V). Then, the charge pump is disabled around $1ms$ and the output capacitor starts to charge up to 3.3V. Fig. 6.13 shows a zoomed version of this plot. The efficiency plot of the charge pump is shown in Fig. 6.14. The equation that is used for efficiency is:

$$Eff = \frac{-V_{out} \times I_{load}}{V_{DD} \times I_{Vdd}} \tag{6.3}$$

Fig. 6.14 shows that the maximum efficiency is around 11% and around 20μA of load current.

6.1.3 Programming FG Transistors Using Negative Voltages

Figure 6.15 shows our complete programming circuit that permits using below-ground voltages for inducing injection. The left-hand side of Fig. 6.15 is a compact version of our circuit

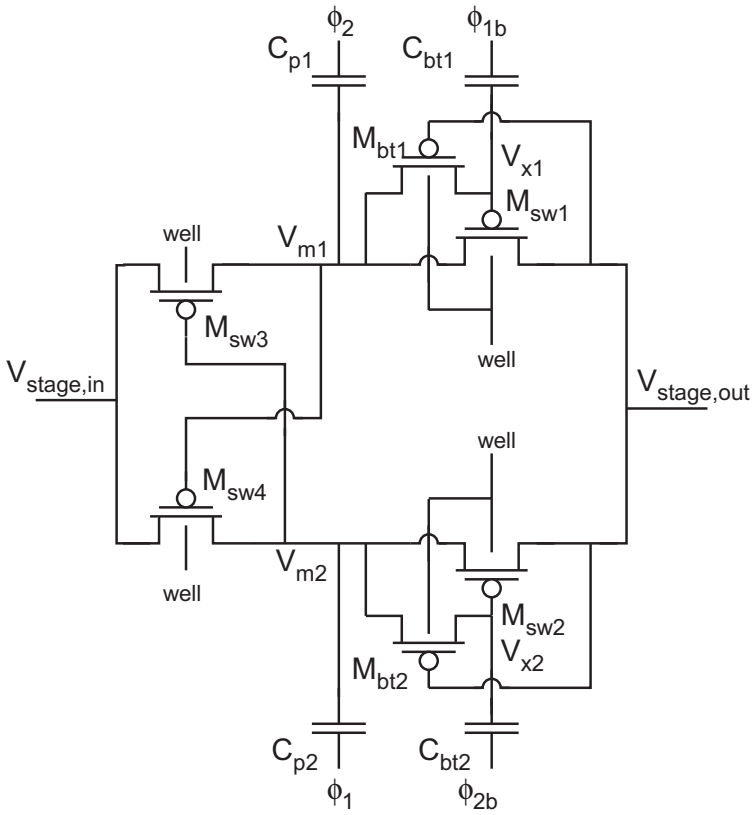


Figure 6.9: Circuit diagram of our charge transfer switches (CTS).

for linearizing injection (Fig. 6.1(b)); a common-source amplifier (M_1-I_2) replaces the high-gain inverting amplifier in Fig. 6.1(b), and M_2 acts as the current source I_{ref}. Additionally, an operational transconductance amplifier (OTA) is connected in a feedback loop to stop injection when V_{cg} reaches a desired target voltage, V_{targ}. Transistors M_{Reset} and M_{bias} are used for initiating injection, as will be discussed shortly.

To program the FG to a desired target, the circuit is placed in "program mode," which is done by setting "disable" to connect the source of M_{inj} to V_s and by setting "select" to connect the source of $M_{circuit}$ to ground. When the negative charge pump is enabled, there will be a large V_{sd} across M_{inj} causing injection currents, but the V_{sd} across $M_{circuit}$ will be too small to cause any injection. Assuming that the FG has been tunneled prior to injecting, node V_{cg} will ramp linearly up as electrons are injected onto the FG via injection through M_{inj}. While V_{cg} is well below the desired V_{targ}, the OTA will supply a constant current equal to its bias current, which is then rectified and mirrored to M_{inj} through M_2-M_3. V_{fg} and V_s stay constant while injecting due to the feedback structure provided by M_1-I_2. This process continues until V_{cg} gets close to V_{targ}, and the output current of the OTA decreases, which decreases the bias current of M_{inj}. As in [2], this reduction in the current through M_{inj} reduces the programming rate as the target is reached, thereby resulting in a better tradeoff between programming speed and accuracy. At this point, M_1 goes into the deep triode region, V_{cg} jumps up to the supply voltage, and injection is completely

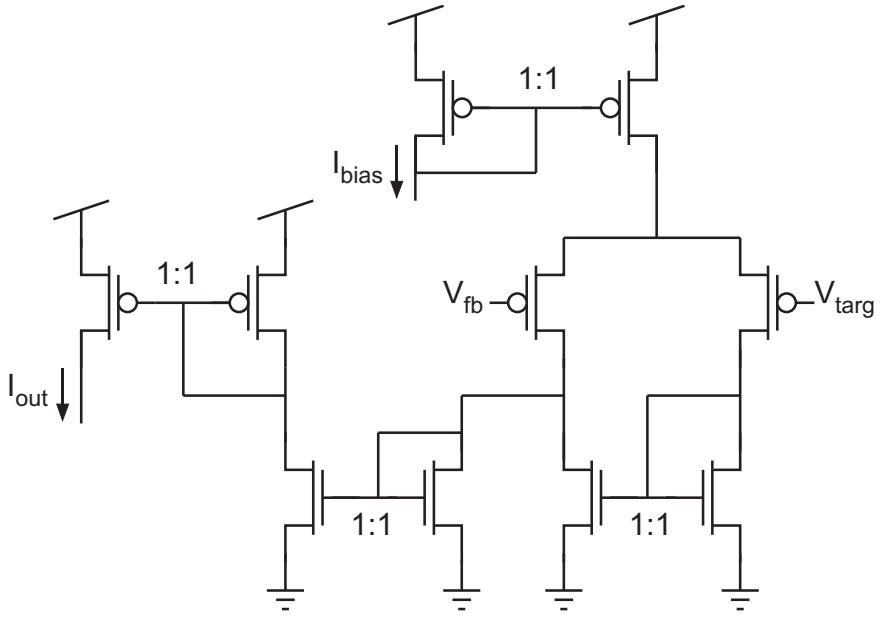


Figure 6.10: Circuit diagram of the OTA used in negative charge pump.

shut off.

After being programmed up to the desired target, the circuit can be placed in a "read mode;" the charge pump is disabled which sends $V_{negative}$ to V_{DD}, and the two FG transistors are completely disconnected from the OTA and the other transistors of Fig. 6.15. A constant voltage is placed on V_{cg}, and "select" connects the source of $M_{circuit}$ to V_{DD}. The readout current, $I_{circuit}$, is connected to whatever circuit it is biasing, and it provides a constant and precise current source. We also use the same structure as read mode for the tunneling erasure process, except V_{tun} is elevated to approximately 15.5V and V_{cg} is set to ground.

When first placing a selected FG cell into program mode, V_{cg} frequently settles to a high voltage due to pre-existing biasing conditions from a previous read-mode setting or from being a non-selected device. Consequently, injection will never start because $V_{cg} > V_{targ}$, thereby shutting off the current in M_2-M_3. Transistors M_{Reset} and M_{bias} have been added to provide a short pulse to "reset" the feedback loop, and cause V_{cg} to go to its other stable point, which is $< V_{targ}$. M_{bias} is optional and can be used to limit the current during this short pulse.

Figure 6.16 shows a timing diagram of the programming process. Before starting injection, the memory cell is in read mode, so V_{cg} is at a fixed voltage, and the FG has been tunneled. When the injection process starts, V_{cg} starts from ground and ramps up quickly. When injection is finished, the charge pump is disabled and the memory cell is configured in read mode, again. The charge pump can remain on after the injection process is finished. This will not cause any further injection after V_{cg} has tripped to V_{DD}.

Figure 6.17 shows the performance of the memory cell and programmer combination. We programmed the memory cell to 21 different target values (V_{targ}) spaced evenly from 1.2V to 3.2V,

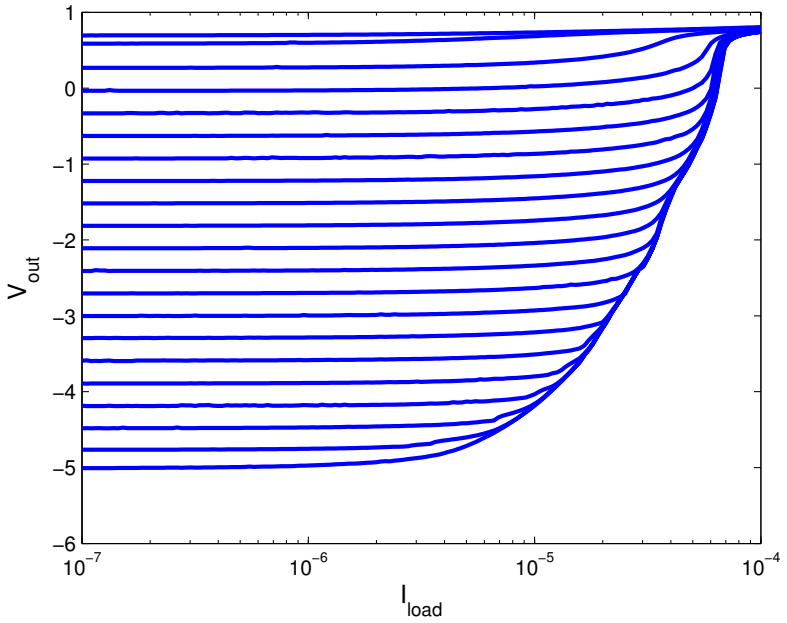


Figure 6.11: Measured V_{out} vs. I_{load} of the charge-pump for load currents from 100nA to 100μA

and then we read the current through the $M_{circuit}$ in read mode. This process was repeated 100 times for each target voltage to check the repeatability of the programming process. Figure 6.17 (top) shows the average measured currents in read mode for different target voltages. Figure 6.17 (bottom) shows that the proposed structure can program currents as low as 1nA with a standard deviation $\leq 1.3\%$.

The goal of designing this memory cell is to use it in a large array of FG transistors. Therefore, we need to ensure that we can select a specific FG transistor when injecting and disable the rest of the FG transistors. To disable a specific FG transistor, we set "disable" such that the source terminal of the unselected M_{inj} transistors connects to ground (see Fig. 6.2). Then, even when the charge pump is enabled, the V_{sd} of the disabled transistors will be small and will not cause unwanted injection in the disabled memory cell.

6.2 High-Voltage Negative Charge-Pumps in 0.35μm CMOS process

6.2.1 First Negative Charge-Pump

In this section, two negative charge-pumps in a 0.35μm standard CMOS process are presented. The circuit schematic of the first negative charge pump is presented in Fig. 6.18. A similar design was presented in section 6.1.2. However, there are a few structural differences between these two designs. First, the proposed charge-pump in this section is designed in a 0.35μm CMOS process. Second, the number of charge-transfer switches in this design is 6. This is because the supply

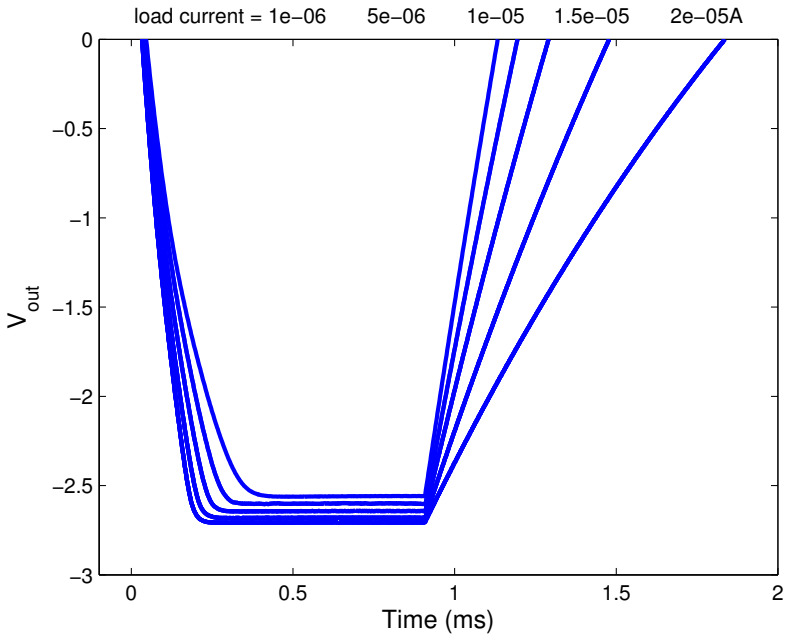


Figure 6.12: Transient response of the negative charge pump

voltage used in the 0.35μm CMOS process is smaller than the 0.5μm CMOS process. Thus, the voltage shift created in each stage is lower in 0.35μm CMOS process and therefore we will need a higher number of stages to generate the same voltage. Third, in this design, the Edgifier circuit is not used. This is because the operating frequency of the proposed charge-pump is higher than the operational range of the Edgifier circuit. The same exact charge-transfer switches are used in both designs (Fig. 6.19).

This operation of this circuit is exactly like the charge pump presented in the previous section. The charge-pump stages generate a voltage that is close to $-NV_{DD}$. Due to the effect of the load current, the generated output voltage would be a few milli-volts above the target negative voltage. The resistor divider will generate a positive voltage (V_{fb}). This feedback voltage is compared with the target voltage (V_{targ}) and a proportional current is generated at the output of the OTA. The generated current will be fed to the ring oscillator and a proportional clock frequency is generated at the output of the ring oscillator. Finally, the non-overlapping clock generator will generate four non-overlapped clock signals to drive the charge-transfer switches.

6.2.2 Experimental results

The proposed circuit is fabricated in a 0.35μm standard CMOS technology. The supply voltage of the proposed circuit is at $2.5V$ and the bias current of the OTA is 1μA. The measured load regulation of the proposed charge pump is shown in Fig. 6.20. The target voltage of the proposed circuit was varied from $0V$ to $1V$ and the charge-pump generates negative voltages as low as $-5V$.

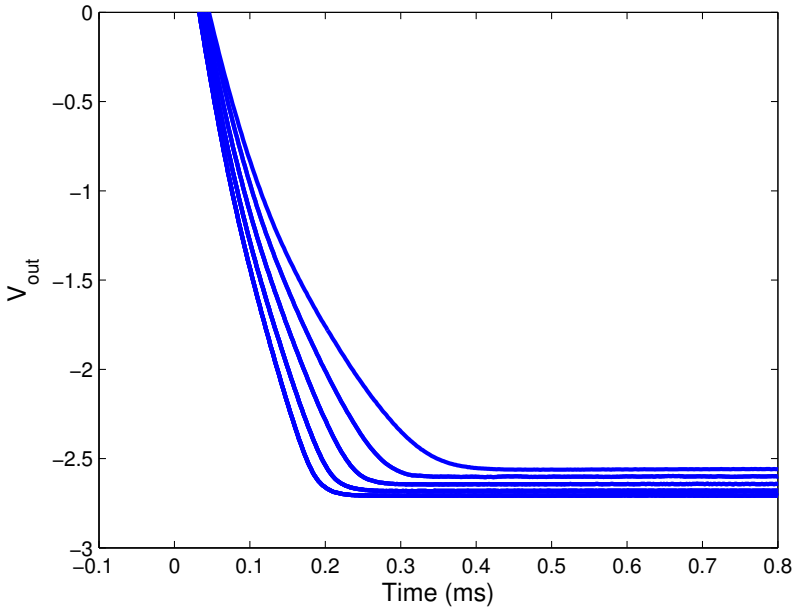


Figure 6.13: Falling edge of the negative charge pump output

6.2.3 Second Negative Charge-Pump

The charge pump that we presented in the previous section suffers from voltage gain. Ideally, the output of an N stage negative Dickson charge pump is $-NV_{DD}$ in steady state. However, in reality, when a diode-connected transistor turns on, it goes into saturation region and the voltage drop across the transistor is V_{th}. The threshold voltage of the PMOS transistors across the diode-connected transistor chain increases from the input to the output due to the body effect. This is because the body of PMOS transistors are connected to ground and the body-source voltage of the transistors keep increasing as it gets closer to the final stages. The voltage gain at each stage is reduced by this voltage drop across the charge transfer transistors. Accordingly, increasing the number of stages will not help to generate higher output voltages. Simulation results of the first negative charge pump show that the maximum achievable negative voltage in a 0.35μm CMOS process is approximately $-5V$.

In order to improve the output voltage range and the load current range of the charge pump, a second charge pump is presented in a standard 0.35μm CMOS process. The circuit schematic of the proposed charge pump is shown in Fig. 6.21(a). There are two major differences between this charge pump and the charge pump presented in the previous section. First, lower number of charge transfer switches are used in this design. Second, four clock boosting circuits are added in between the non-overlapping clock generator and the charge transfer switches. The structure of the charge transfer switch is shown in Fig. 6.21(c). The clock booster circuits receive four non-overlapped clocks from the non-overlapping clock generator and generate four non-overlapped clocks with larger high logic levels. Thus the effective voltage shift per switching stage will be

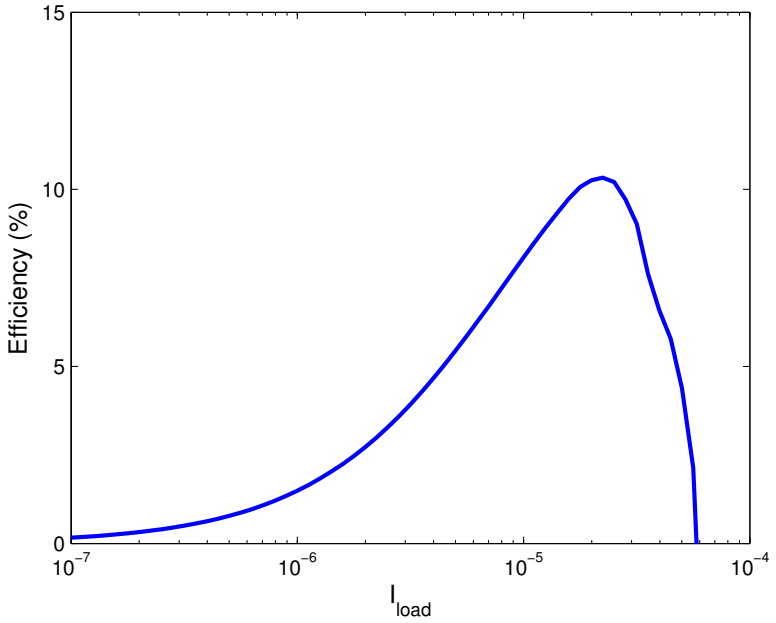


Figure 6.14: Power Efficiency of the negative charge pump

improved. Thus a larger negative voltage is possible using this structure. The circuit schematic of the clock booster is shown in Fig. 6.21(b). Assuming that V_{in} is high, the top plate of C_1 shifts up by one V_{DD} and the top plate of C_2 stays at the same level. M_1 and M_2 have a cross-coupled structure. Thus, M_1 turns off and M_2 turns on. This will pull the top plate of C_2 to a higher level (HV_{DD}). Since the gate of M_4 was high, this transistor will be on and the gate of M_3 is at ground level. Thus M_3 is on and V_{out} is fixed at a higher voltage (HV_{DD}).

6.2.4 Experimental results

The proposed circuit is fabricated in a 0.35μm standard CMOS technology. The supply voltage of the proposed circuit is at $2.5V$ and the bias current of the OTA is 10μA. The measured load regulation of the proposed charge pump is presented in Fig. 6.22. The target voltage of the proposed structure was swept from $0.5V$ to $1.5V$. This design can generate $-5.5V$ output under higher load currents.

6.3 Conclusion

We have presented a compact floating-gate memory cell. Indirect programming of the FG transistors made it possible to use negative voltages to program the FG transistor and circumvent the issue of using selection circuits operating below ground. Two closed-loop regulating charge pumps were also presented to generate the negative voltages required for programming.

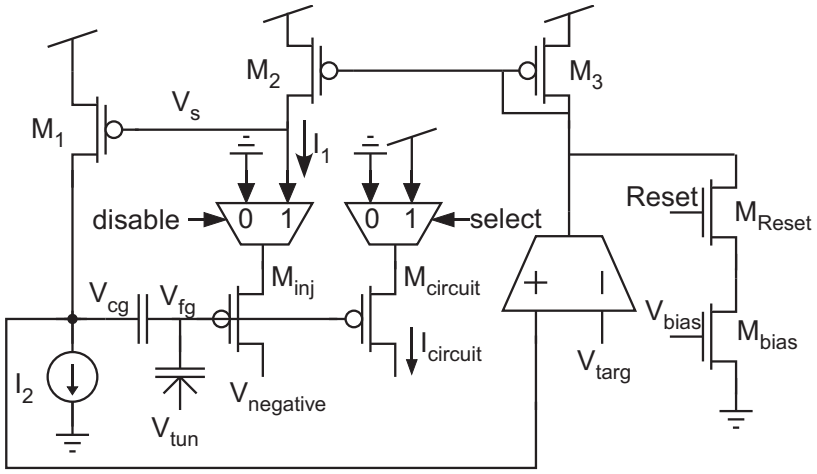


Figure 6.15: Complete below-ground programming structure.

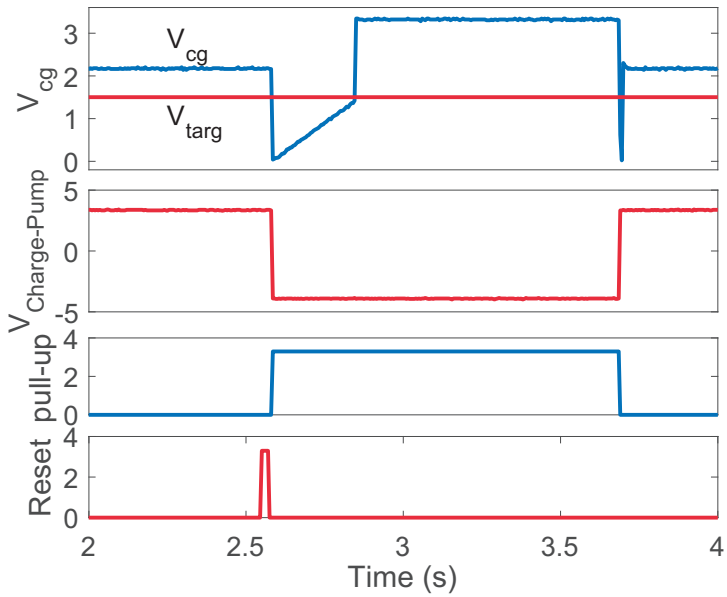


Figure 6.16: Timing diagram of injecting an FG transistor to a desired value ($V_{targ} = 1.5V$), showing the control gate voltage, the output of the charge pump, pull-up voltage, and Reset voltage.

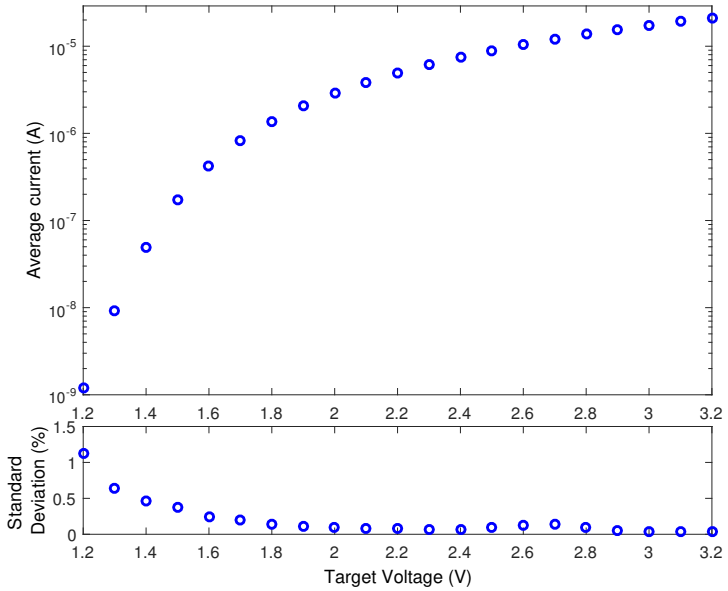


Figure 6.17: (top) Average programmed current through $M_{circuit}$ in read mode over 100 programming iterations, for multiple V_{targ} values. (bottom) Standard deviation of the programmed current.

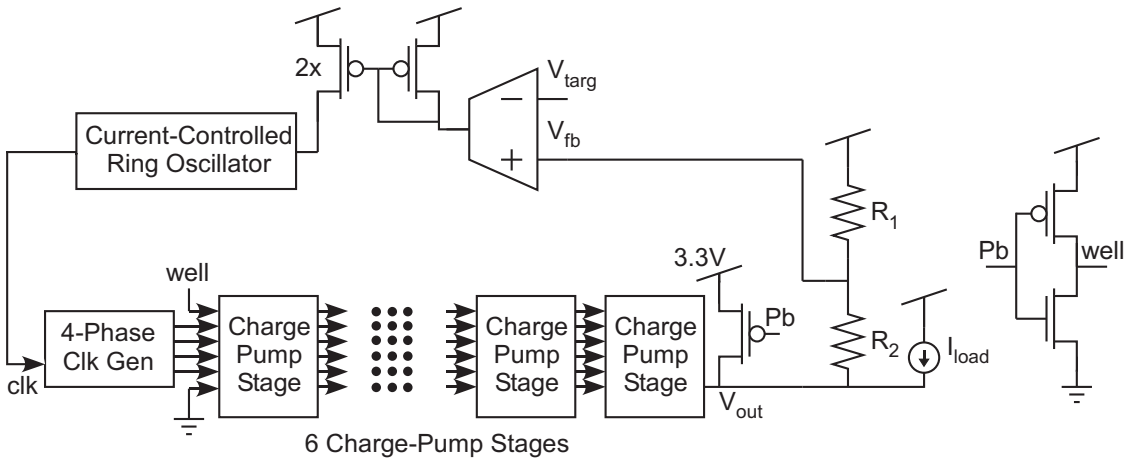


Figure 6.18: Circuit structure of the first proposed charge pump

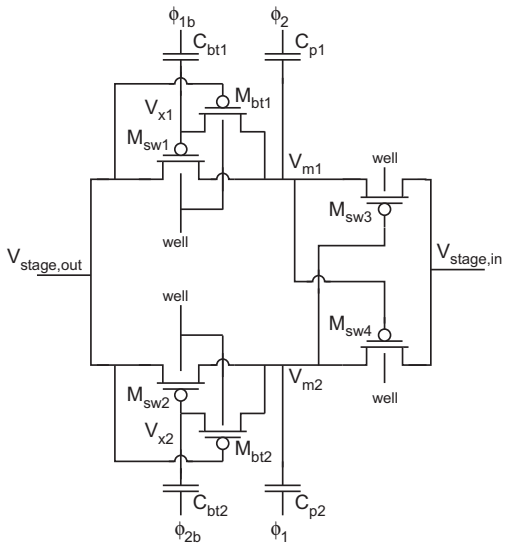


Figure 6.19: Circuit structure of the charge-transfer switches

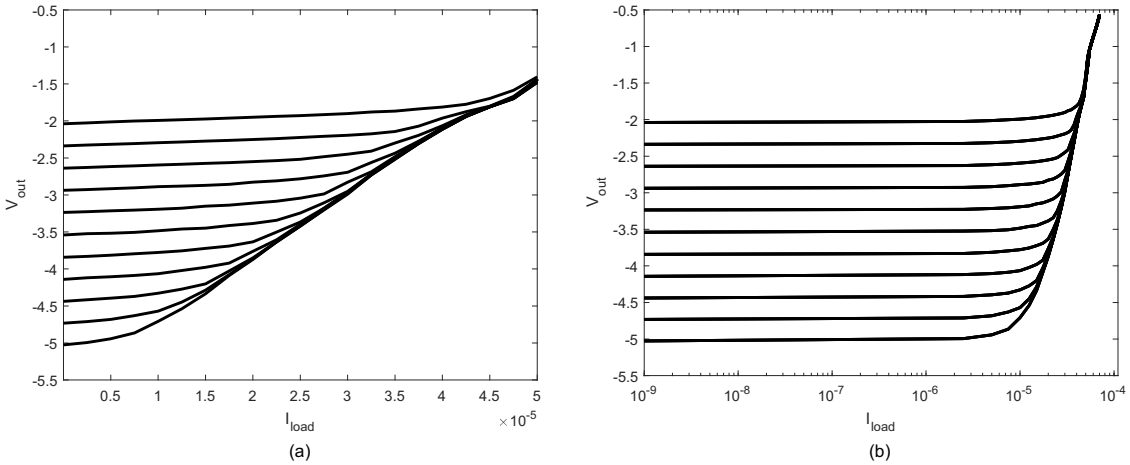


Figure 6.20: Measured load regulation of the first negative charge-pump

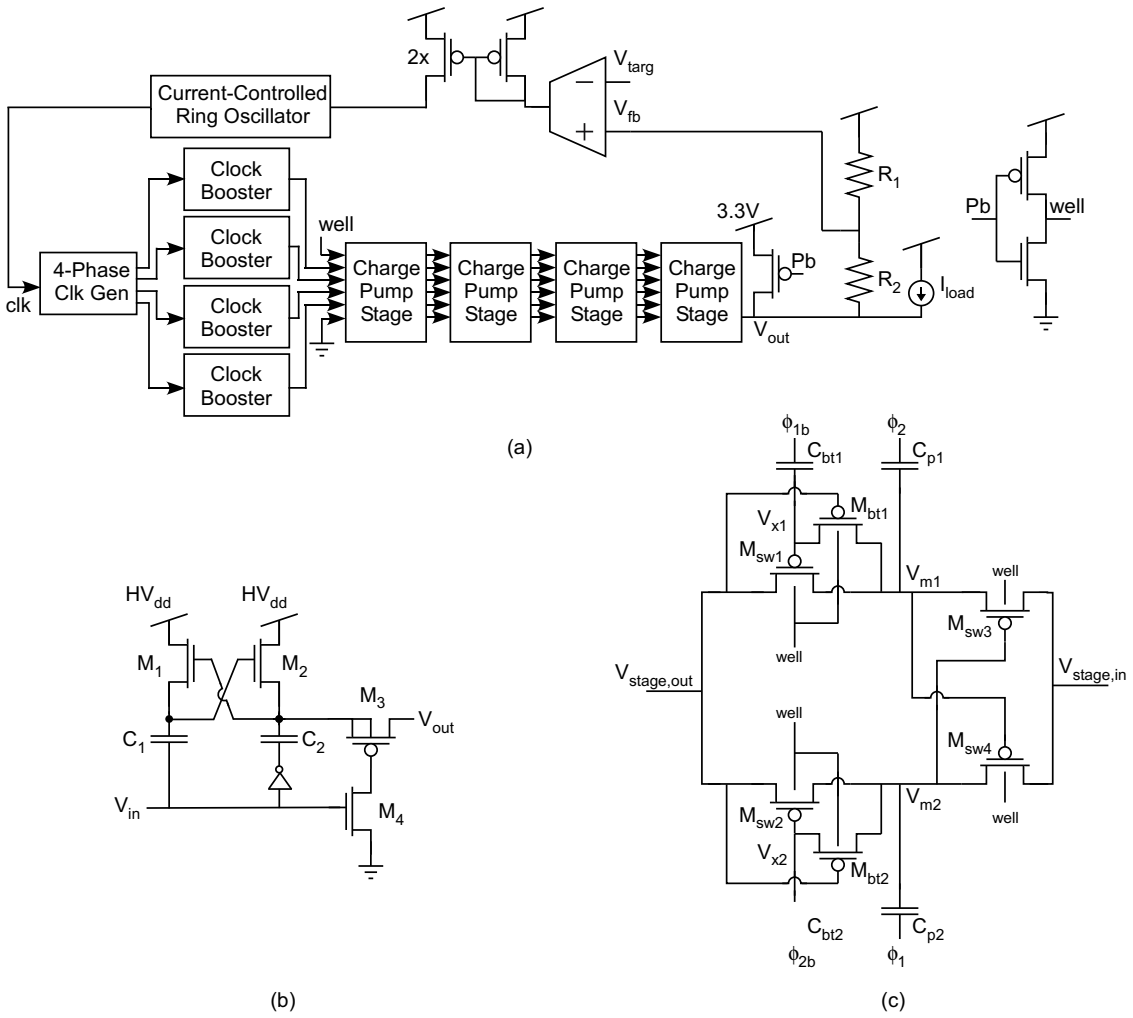


Figure 6.21: Circuit structure of the second proposed charge pump

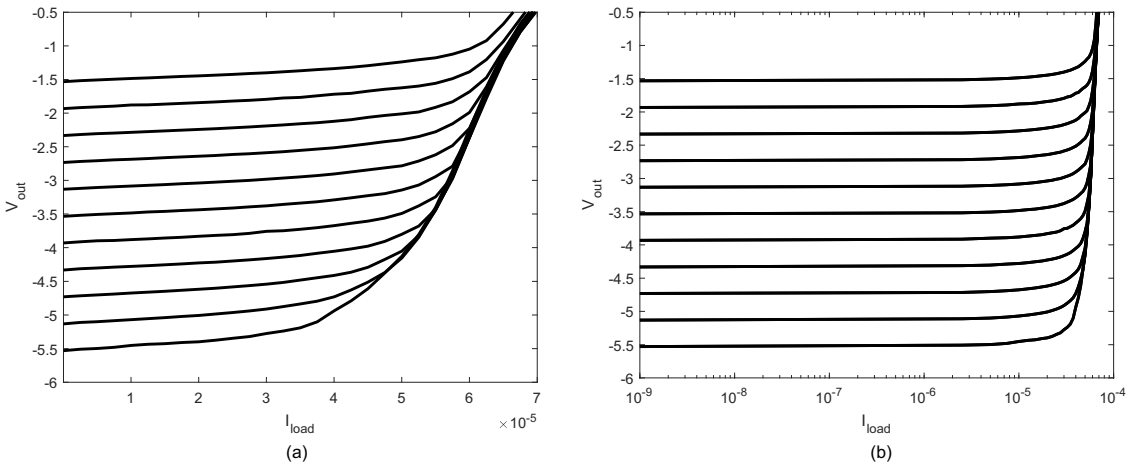


Figure 6.22: Measured load regulation of the second negative charge-pump

Chapter 7

Conclusions and Future Work

The concept of the Internet of Things (IoT) is currently a highly significant topic. IoT is the network of devices that we use in our daily life and other devices that include sensors and actuators which enables these devices to communicate and exchange data. These networks of devices share data among sensor nodes that are sent to a distributed system for the analysis of the data.

The benefits of reconfigurable analog signal processing make Field Programmable Analog Arrays (FPAAs) attractive for resource-constrained sensing applications. Unfortunately, the cost of dense analog data storage, which includes high power consumption to update the stored analog data or high infrastructure overhead to write nonvolatile storage, limits the use of large-scale FPAAs in low-power systems.

In order to make large-scale programmable analog systems, low-power programming circuits are required. We showed that to change the stored analog data, Fowler-Nordheim tunneling and hot-electron injection are used. Both of these techniques require highly stable and low ripple voltages. We must use step-up voltage converters for these two procedures. Thus, we presented the design procedure and experimental results of two charge pump circuits for FG programming applications.

Any System-on-Chip (SoC) will require supporting circuits to provide stable supply and mid-rail voltages. Voltage reference cells are used to provide a stable voltage across a wide range of temperature and supply variations. We presented the design procedure and experimental results of two voltage reference cells that can be used in SoC applications. We also present the experimental results of two voltage regulators which generate V_{DD} and V_{mid} for analog and digital circuits in our system.

The final goal of this work is to present a new programming technique for analog applications namely below ground programming of FG transistors. Negative voltages are used to program the FG transistors in digital applications. However, programming analog data in FG transistors using negative voltages is a challenging task especially in standard CMOS processes. We presented a technique to program the FG transistor using the indirect method of programming. Also, we presented a technique to select a target FG transistor in an array of FG transistors for large-scale

applications. We proved that this technique is possible in standard CMOS processes. Also, two charge-pump structures are presented which can generate negative voltages and provide enough load currents for the below ground programming technique.

7.1 Future work

We can make improvements to a few of the proposed designs. Improving the load current range of the injection charge pump and the negative charge pump will reduce the overall complexity of the full FPAA system. The FPAA system has other circuits that require high supply voltages during the programming of a target FG transistor. For example, the injection charge pump can be used to provide the voltages for other sub-circuits like Digital-to-Analog converters (DACs) and serial peripheral interfaces (SPIs). These circuits may consume high power in transients to charge the parasitic capacitors. Thus, improving the load current range of the injection charge pump will circumvent the requirement of designing multiple high-voltage step-up converters which reduces both the power consumption and the die area of the overall FPAA system.

The voltage reference cells and the supply voltage regulator presented in this dissertation use on-chip resistors. These resistors generate thermal noise which degrades the noise performance of these circuits. Besides, large resistors employed in these designs will occupy big die area which increases the cost of the presented integrated circuits. Thus, a resistor-less design can reduce the cost of the design.

The curvature compensation circuit presented in this dissertation can be employed at the output stage of any voltage reference cell. Besides, improving the process sensitivity of the presented voltage reference cell makes this design a good candidate for SoC applications. The mid-rail voltage regulator presented in this dissertation has $8mV$ offset across -10μA to 10μA of load currents. This offset can be improved by using high-gain amplifiers inside the regulation loops.

The below-ground programming technique was checked for a two-by-one array of FG transistors. However, trying this idea for a big array of FG transistors would be beneficial for future large-scale applications like FPAAs. It must be checked if programming a target FG memory cell would have any effects on the charge programmed onto the neighbor FG transistors in the array. Shift registers and other digital circuits are also required in the design of a big array of FG memory cells.

All of the presented circuits must be incorporated into an FPAA system. Employing these circuits into a larger system helps to understand the requirements of these circuits realistically. This is particularly challenging for an FPAA using the below-ground programming technique.

All of the circuits presented in this dissertation have other potential applications too. The high-voltage charge pumps can be used in MEMS devices, electret earphones, and mobiles DRAMs. The voltage regulators and the voltage reference cells can be used in implantable biomedical microsystems and data converters. The proposed negative charge pumps can be employed in TFT LCD

display drivers and high impedance microstimulation. These designs can be incorporated into the mentioned applications.

References

[1] C. Huang, P. Sarkar, and S. Chakrabartty, "Rail-to-rail, linear hot-electron injection programming of floating-gate voltage bias generators at 13-bit resolution," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2685–2692, Nov., 2011.

[2] B. Rumberg and D. W. Graham, "A floating-gate memory cell for continuous-time programming," in *Proceedings of the Midwest Symposium on Circuits and Systems*, Aug., 2012, pp. 214–217.

[3] B. Rumberg, D. W. Graham, and M. M. Navidi, "A regulated charge pump for tunneling floating-gate transistors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 3, pp. 516–527, March, 2017.

[4] A. Umezawa, S. Atsumi, M. Kuriyama, H. Banba, K. Imamiya, K. Naruke, S. Yamada, E. Obi, M. Oshikiri, T. Suzuki, and S. Tanaka, "A 5-V-only operation 0.6μm flash EEPROM with row decoder scheme in triple-well structure," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 11, pp. 1540–1546, Nov., 1992.

[5] O. Wong, H. Wong, W. Tam, and C. Kok, "A comparative study of charge pumping circuits for flash memory applications," *Microelectronics Reliability*, vol. 52, no. 4, pp. 670 – 687, 2012.

[6] M. Bloch, C. Lauterbauch, and W. Weber, "High efficiency charge pump circuit for negative high voltage generation at 2V supply voltage," in *Proceedings of the European Solid-State Circuits Conference*, Sept., 1998, pp. 100–103.

[7] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells-an overview," *in Proceedings of the IEEE*, vol. 85, no. 8, pp. 1248–1271, Aug., 1997.

[8] E. Suzuki, H. Hiraishi, K. Ishii, and Y. Hayashi, "A low-voltage alterable EEPROM with Metal-Oxide-Nitride-Oxide-Semiconductor ($MONOS$) structures," *IEEE Transactions on Electron Devices*, vol. 30, no. 2, pp. 122–128, Feb., 1983.

[9] E. Suzuki, K. Miura, Y. Hayasi, R. P. Tsay, and D. K. Schroder, "Hole and electron current transport in metal-oxide-nitride-oxide-silicon memory structures," *IEEE Transactions on Electron Devices*, vol. 36, no. 6, pp. 1145–1149, June, 1989.

[10] B. Rumberg and D. Graham, "Efficiency and reliability of Fowler-Nordheim tunnelling in $CMOS$ floating-gate transistors," *Electronics Letters*, vol. 49, no. 23, pp. 1484–1486, Nov., 2013.

[11] M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *IEEE Transactions on Electron Devices*, vol. 15, no. 9, p. 686, Sept., 1968.

[12] L. Carley, "Trimming analog circuits using floating-gate analog MOS memory," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 6, pp. 1569–1575, Dec., 1989.

[13] International Technology Roadmap for Semiconductors, *Process Integration, Devices, and Structures*. http://www.itrs.net/Links/2013ITRS/Home2013.htm, 2013.

[14] S. H. Song, K. C. Chun, and C. H. Kim, "A bit-by-bit re-writable Eflash in a generic 65 nm logic process for moderate-density nonvolatile memory applications," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 8, pp. 1861–1871, Aug., 2014.

[15] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. VLSI Signal Process. Syst. Signal, Image Video Technology*, vol. 8, no. 1, pp. 9–25, 1994.

[16] *Predictive Technology Model*. http://ptm.asu.edu, Aug., 2015.

[17] S. Shapero, C. Rozell, A. Balavoine, and P. Hasler, "A scalable implementation of sparse approximation on a field programmable analog array," in *Proceedings of the Biomedical Circuits and Systems Conference*, Nov., 2011, pp. 141–144.

[18] B. Rumberg, D. W. Graham, S. Clites, B. M. Kelly, M. M. Navidi, A. Dilello, and V. Kulathumani, "Ramp: Accelerating wireless sensor hardware design with a reconfigurable analog/mixed-signal platform," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, ser. IPSN '15, 2015, pp. 47–58.

[19] A. Bandyopadhyay, G. Serrano, and P. Hasler, "Adaptive algorithm using hot-electron injection for programming analog computational memory elements within 0.2% of accuracy over 3.5 decades," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 9, pp. 2107–2114, Sept., 2006.

[20] C. Diorio, "A p-channel MOS synapse transistor with self-convergent memory writes," *IEEE Transaction on Electron Devices*, vol. 47, no. 2, pp. 464–472, Feb., 2000.

[21] J. Chung, M. Jeng, G. May, P. Ko, and C. Hu, "Hot-electron currents in deep-submicrometer MOSFETs," in *Tech. Dig. Int. Electron Devices Meeting*, San Francisco, CA, Dec., 1988, pp. 200–203.

[22] T.-C. Ong, P.-K. Ko, and C. Hu, "Hot-carrier current modeling and device degradation in surface-channel p-MOSFETs," *IEEE Trans. Electron Devices*, vol. 37, no. 7, pp. 1658–1666, July, 1990.

[23] S. Chakrabartty and G. Cauwenberghs, "Fixed-current method for programming large floating-gate arrays," in *IEEE International Symposium on Circuits and Systems*, vol. 4, May, 2005, pp. 3934–3937.

[24] C. G. O. P. Micheloni, Rino, *Memories in Wireless Systems*. Springer, 2008.

[25] Micron Semi., *NAND Flash 101: An Introduction to NAND Flash and How to Design It In to Your Next Product*. https://www.micron.com/resource-details/fea5cfd9-ee93-47f4-b2af-cd494d3291c3, 206.

[26] B. Rumberg and D. Graham, "A low-power and high-precision programmable analog filter bank," *IEEE Trans. Circuits Syst. II*, vol. 59, no. 4, pp. 234–238, April, 2012.

[27] W. Borchani, K. Aono, N. Lajnef, and S. Chakrabartty, "Monitoring of postoperative bone healing using smart trauma-fixation device with integrated self-powered piezo-floating-gate sensors," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1463–1472, July, 2016.

[28] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with Floating-Gate storage in 0.13μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan., 2015.

[29] Y. L. Wong, M. Cohen, and P. Abshire, "On-line histogram equalization for Flash adc," in *Proceedings of the International Symposium on Circuits and Systems*, May, 2007, pp. 3598–3601.

[30] Y. L. Wong, M. H. Cohen, and P. A. Abshire, "A 1.2 GHz adaptive floating gate comparator with 13-bit resolution," in *Proceedings of the International Symposium on Circuits and Systems*, May, 2005, pp. 6146–6149 Vol. 6.

[31] V. Srinivasan, D. Graham, and P. Hasler, "Floating-gate transistors for precision analog circuit design: An overview," in *Proceedings of the IEEE Midwest Symp. Circuits Syst.*, vol. 1, Covington, KY, Aug., 2005, pp. 71–74.

[32] B. Rumberg and D. Graham, "A low-power field-programmable analog array for wireless sensing," in *Proc. Int. Symp. Quality Electronic Design*, Santa Clara, CA, March, 2015, pp. 542–546.

[33] J. Dickson, "On-chip high-voltage generation in MNOS integrated circuits using an improved voltage multiplier technique," *IEEE J. Solid-State Circuits*, vol. 11, no. 3, pp. 374–378, June, 1976.

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

[35] E. Racapé and J.-M. Daga, "A PMOS-switch based charge pump, allowing lost cost implementation on a CMOS standard process," in *Proc. 31st European Solid-State Circuits Conf.*, Sept., 2005, pp. 77–80.

[36] N. Li, Z. Huang, M. Jiang, and Y. Inoue, "High efficiency four-phase all PMOS charge pump without body effect," in *Proc. Int. Conf. Commun., Circuits Syst.*, May, 2008, pp. 1083–1087.

[37] J. Shin, I. Chung, Y. Park, and H. Min, "A new charge pump without degradation in threshold voltage due to body effect," *IEEE J. Solid-State Circuits*, vol. 35, no. 8, pp. 1227–1230, Aug., 2000.

[38] B. Gregoire, "A compact switched-capacitor regulated charge pump power supply," *IEEE J. Solid-State Circuits*, vol. 41, no. 8, pp. 1944–1953, Aug., 2006.

[39] C.-H. Wu and C.-L. Chen, "A low-ripple charge pump with continuous pumping current control," in *Proc. IEEE Midwest Symp. Circuits Syst.*, Aug., 2008, pp. 722–725.

[40] Y. H. Kang, J.-K. Kim, S. W. Hwang, J. Y. Kwak, J.-Y. Park, D. Kim, C. H. Kim, J. Y. Park, Y.-T. Jeong, J. N. Baek, S. C. Jeon, P. Jang, S.-H. Lee, Y.-S. Lee, M.-S. Kim, J.-Y. Lee, and Y.-H. Choi, "High-voltage analog system for a mobile NAND flash," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 507–517, Feb., 2008.

[41] E. Bayer and H. Schmeller, "Charge pump with active cycle regulation—closing the gap between linear and skip modes," in *Proc. IEEE Power Electron. Specialists Conf.*, vol. 3, 2000, pp. 1497–1502.

[42] J.-Y. Lee, S.-E. Kim, S.-J. Song, J.-K. Kim, S. Kim, and H.-J. Yoo, "A regulated charge pump with small ripple voltage and fast start-up," *IEEE J. Solid-State Circuits*, vol. 41, no. 2, pp. 425–432, Feb., 2006.

[43] C. Mingyang, W. Xiaobo, and Z. Menglian, "Novel high efficiency low ripple charge pump using variable frequency modulation," in *Proc. Int. Conf. Microelectronics*, Dec., 2010, pp. 228–231.

[44] L. Aaltonen and K. Halonen, "On-chip charge-pump with continuous frequency regulation for precision high-voltage generation," in *Ph.D. Research in Microelectronics and Electronics (PRIME)*, San Francisco, CA, July, 2009, pp. 68–71.

[45] C. Yoo, "A CMOS buffer without short-circuit power consumption," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 47, no. 9, pp. 935–937, Sept., 2000.

[46] C. Lee and P. Mok, "A monolithic current-mode CMOS DC-DC converter with on-chip current-sensing technique," *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 3–14, Jan., 2004.

[47] R. Barnett and J. Liu, "An EEPROM programming controller for passive UHF RFID transponders with gated clock regulation loop and current surge control," *IEEE J. Solid-State Circuits*, vol. 43, no. 8, pp. 1808–1815, Aug., 2008.

[48] K.-Y. Kim, Y. Kim, D.-C. Lee, Y.-R. Kang, H. Kim, S.-W. Kim, and J. Park, "An energy efficient V_{PP} generator with fast ramp-up time for mobile DRAM," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1488–1494, June, 2011.

[49] J. H. Tsai, C. Y. Tseng, W. K. Tseng, T. K. Shia, and P. C. Huang, "An integrated 12-V electret earphone driver with symmetric cockcroft-walton pumping topology for in-ear hearing aids," in *Proc. IEEE Asian Solid State Circuits Conf.*, Nov., 2012, pp. 45–48.

[50] J. H. Tsai, S. A. Ko, C. W. Wang, Y. C. Yen, H. H. Wang, P. C. Huang, P. H. Lan, and M. H. Shen, "A 1V input, 3 V-to-6 V output, 58%-efficient integrated charge pump with a hybrid topology for area reduction and an improved efficiency by using parasitics," *IEEE J. Solid-State Circuits*, vol. 50, no. 11, pp. 2533–2548, Nov., 2015.

[51] C. Y. Tseng, S. C. Chen, T. K. Shia, and P. C. Huang, "An integrated 1.2V-to-6V CMOS charge-pump for electret earphone," in *Proc. IEEE Symp. VLSI Circuits*, June, 2007, pp. 102–103.

[52] K.-H. Lee, S.-C. Wang, and Y.-C. King, "Self-convergent scheme for logic-process-based multilevel/analog memory," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2676–2681, Dec., 2005.

[53] P. D. Smith, M. Kucic, and P. Hasler, "Accurate programming of analog floating-gate arrays," in *IEEE International Symposium on Circuits and Systems*, vol. 5, 2002, pp. V–489–V–492.

[54] K. Wong and D. Evans, "A 150mA low noise, high PSRR low-dropout linear regulator in 0.13μm technology for RF SoC applications," in *European Solid-State Circuits Conference*, Sept., 2006, pp. 532–535.

[55] A. Amer and E. Sanchez-Sinencio, "A 140mA 90nm CMOS low drop-out regulator with -56db power supply rejection at 10MHz," in *Proceedings of Custom Integrated Circuits Conference (CICC)*, Sept., 2010, pp. 1–4.

[56] C. K. Chava and J. Silva-Martinez, "A frequency compensation scheme for LDO voltage regulators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 6, pp. 1041–1050, June, 2004.

[57] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low voltage and low-current applications," *Analog Integrat. Circuits Signal Process.*, vol. 8, pp. 83–114, 1995.

[58] W. Yan, W. Li, and R. Liu, "Nanopower CMOS sub-bandgap reference with 11 ppm/C temperature coefficient," *Electronics Letters*, vol. 45, no. 12, pp. 627–629, June, 2009.

[59] T. Hirose, K. Ueno, N. Kuroki, and M. Numa, "A CMOS bandgap and sub-bandgap voltage reference circuits for nanowatt power LSIs," in *2010 IEEE Asian Solid-State Circuits Conference*, Nov., 2010, pp. 1–4.

[60] C. Enz and E. Vittoz, "CMOS low-power analog circuit design," in *Emerging Technologies: Designing Low Power Digital Systems*, 1996, pp. 79–133.

[61] Y. Chiang and S. Liu, "Nanopower CMOS relaxation oscillators with sub-100 $ppm/^{\circ}C$ temperature coefficient," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 61, no. 9, pp. 661–665, Sept., 2014.

[62] B. Chatterjee, N. Modak, A. Amaravati, D. Mistry, D. Das, and M. Baghini, "A sub-1V, 120 nW, PVT-variation tolerant, tunable, and scalable voltage reference with 60-dB PSNA," *IEEE Trans. Nanotechnol.*, vol. 16, no. 3, pp. 406–410, May, 2017.

[63] Y. Tsividis, "Operation and modeling of the MOS transistor," *New York: McGraw-Hill*, 1987.

[64] I. Lee, D. Sylvester, and D. Blaauw, "A subthreshold voltage reference with scalable output voltage for low-power IoT systems," *IEEE J. Solid-State Circuits*, vol. 52, no. 5, pp. 1443–1449, May, 2017.

[65] Y. Wang, Z. Zhu, J. Yao, and Y. Yang, "A 0.45V, 15.6nW MOSFET-only sub-threshold voltage reference with no amplifiers," in *Proceedings of the IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov., 2015, pp. 1–4.

[66] K. N. Leung and P. K. T. Mok, "A CMOS voltage reference based on weighted ΔV_{GS} for CMOS low-dropout linear regulators," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 1, pp. 146–150, Jan., 2003.

[67] S. I. Chang, K. AlAshmouny, and E. Yoon, "A 1.5V 120nW CMOS programmable monolithic reference generator for wireless implantable system," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug., 2011, pp. 2981–2984.

[68] Y. Osaki, T. Hirose, N. Kuroki, and M. Numa, "1.2-V supply, 100-nW, 1.09-V bandgap and 0.7-V supply, 52.5-nW, 0.55-V sub-bandgap reference circuits for nanowatt CMOS LSIs," *IEEE J. Solid-State Circuits*, vol. 48, no. 6, pp. 1530–1538, June, 2013.

[69] A. Becker-Gomez, T. L. Viswanathan, and T. R. Viswanathan, "A low-supply-voltage CMOS sub-bandgap reference," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 7, pp. 609–613, July, 2008.

[70] Y. Lam and W. Ki, "CMOS bandgap references with self-biased symmetrically matched current-voltage mirror and extension of sub-1-V design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst*, vol. 18, no. 6, pp. 857–865, June, 2010.

[71] G. A. Rincon-Mora, "Current efficient, low voltage, low drop-out regulators," Ph.D. dissertation, Georgia Institute of Technology, Nov., 1996.

[72] K. N. Leung and Y. S. Ng, "A CMOS low-dropout regulator with a momentarily current-boosting voltage buffer," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 9, pp. 2312–2319, Sept., 2010.

[73] S. Rajapandian, Z. Xu, and K. L. Shepard, "Implicit dc-dc down conversion through charge-recycling," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 846–852, April, 2005.

[74] J. Xiao, A. V. Peterchev, J. Zhang, and S. R. Sanders, "A 4μA quiescent-current dual-mode digitally controlled buck converter IC for cellular phone applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 12, pp. 2342–2348, Dec., 2004.

[75] C. Calligaro, R. Gastaldi, P. Malcovati, and G. Torelli, "Positive and negative CMOS voltage multiplier for 5-V-only Flash memories," in *Proceedings of the Midwest Symposium on Circuits and Systems.*, vol. 1, Aug., 1995, pp. 294–297 vol.1.

[76] M. Gill, R. Cleavelin, S. Lin, I. D'Arrigo, G. Santin, P. Shah, A. Nguyen, J. Esquivel, B. Riemenschneider, and J. Paterson, "A 5-volt contactless array 256 kbit flash EEPROM technology," in *Technical Digest., International Electron Devices Meeting*, Dec., 1988, pp. 428–431.

[77] D. Matousek, "Comparison of selected architectures of negative charge pumps with new design," in *Proceedings of the International Conference Radioelektronika*, April, 2014, pp. 1–4.

[78] T. Yamazoe, H. Ishida, and Y. Nihongi, "A charge pump that generates positive and negative high voltages with low power-supply voltage and low power consumption for non-volatile memories," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May, 2009, pp. 988–991.

[79] M. Huang, L. Okamura, Y. Wang, and T. Yoshihara, "A 1.5V four phase switched polarity charge pump," in *Proceedings of the International Conference on Communications, Circuits and Systems*, July, 2009, pp. 688–692.

[80] T. Tanzawa and T. Tanaka, "A dynamic analysis of the dickson charge pump circuit," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1231–1240, Aug., 1997.

[81] J. C. Chen, T. H. Kuo, L. E. Cleveland, C. K. Chung, N. Leong, Y. K. Kim, T. Akaogi, and Y. Kasa, "A 2.7V only 8Mb×16 NOR flash memory," in *Proceedings of the Symposium on VLSI Circuits. Digest of Technical Papers*, June, 1996, pp. 172–173.

[82] H. Lin, J. Lu, and Y.-T. Lin, "A new 4-phase charge pump without body effects for low supply voltages," in *Proceedings. IEEE Asia-Pacific Conference on ASIC,*, 2002, pp. 53–56.

[83] Y. C. Huang, M. D. Ker, and C. Y. Lin, "Design of negative high voltage generator for biphasic stimulator with SoC integration consideration," in *Proceedings of the IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Nov., 2012, pp. 29–32.

[84] C. H. Wu, "Multi-phase charge pump generating positive and negative high voltages for TFT-LCD gate driving," in *Proceedings of the IEEE International Symposium on Electronic Design, Test and Applications*, Jan., 2008, pp. 179–183.

[85] M. H. Cheng and Z. W. Wu, "Low-power low-voltage reference using peaking current mirror circuit," *Electronics Letters*, vol. 41, no. 10, pp. 572–573, May, 2005.

[86] Z. k. Zhou, P. s. Zhu, Y. Shi, X. Qu, H. y. Wang, X. m. Zhang, S. Qiu, N. Li, C. Gou, Z. Wang, and B. Zhang, "A resistorless CMOS voltage reference based on mutual compensation of V_T and V_{TH}," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 9, pp. 582–586, Sept., 2013.

[87] Y. Zhu, F. Liu, Y. Yang, G. Huang, T. Yin, and H. Yang, "A 115dB PSRR CMOS bandgap reference with a novel voltage self-regulating technique," in *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*, Sept., 2014, pp. 1–4.

[88] S. Lee, H. Lee, J. K. Woo, and S. Kim, "Low-voltage bandgap reference with output-regulated current mirror in 90 nm CMOS," *Electronics Letters*, vol. 46, no. 14, pp. 976–977, July, 2010.

[89] A. I. Kamel, A. Saad, and L. S. Siong, "A high wide band PSRR and fast start-up current mode bandgap reference in 130nm CMOS technology," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, May, 2016, pp. 506–509.

[90] C. Popa, "Power-efficient superior-order curvature corrected voltage reference using CMOS computational circuits," in *International Symposium on Signals, Circuits and Systems, 2005. ISSCS 2005.*, vol. 1, July, 2005, pp. 23–26 Vol. 1.

[91] S.-Y. Chin and C.-Y. Wu, "A new type of curvature-compensated CMOS bandgap voltage references," in *Proceedings of the International Symposium on VLSI Technology, Systems, and Applications*, May, 1991, pp. 398–402.

[92] G. Ge, C. Zhang, G. Hoogzaad, and K. A. A. Makinwa, "A single-trim $CMOS$ bandgap reference with a 3σ inaccuracy of 0.15% from $-40°C$ to $125°C$," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2693–2701, Nov., 2011.

[93] G. D. Vita and G. Iannaccone, "An ultra-low-power, temperature compensated voltage reference generator," in *Proceedings of the IEEE 2005 Custom Integrated Circuits Conference, 2005.*, Sept., 2005, pp. 751–754.

[94] G. Giustolisi and G. Palumbo, "A detailed analysis of power-supply noise attenuation in bandgap voltage references," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 50, no. 2, pp. 185–197, Feb., 2003.

[95] A. E. Buck, C. L. McDonald, S. H. Lewis, and T. R. Viswanathan, "A CMOS bandgap reference without resistors," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 1, pp. 81–83, Jan., 2002.

[96] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, S. Atsumi, and K. Sakui, "A CMOS bandgap reference circuit with sub-1-V operation," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 5, pp. 670–674, May, 1999.

[97] H. Zhang, Z. Xiao, X. Tan, and H. Min, "Low-power sub-1-V compact bandgap reference for passive RFID tags," *Electronics Letters*, vol. 51, no. 11, pp. 815–816, 2015.

[98] A. Kumar, "Trimless second order curvature compensated bandgap reference using diffusion resistor," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, Sept., 2009, pp. 161–164.

[99] A. Becker-Gomez, T. L. Viswanathan, and T. R. Viswanathan, "A low-supply-voltage CMOS sub-bandgap reference," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 55, no. 7, pp. 609–613, July, 2008.

[100] V. V. Ivanov, K. E. Sanborn, and I. M. Filanovsky, "Bandgap voltage references with 1V supply," in *2006 Proceedings of the 32nd European Solid-State Circuits Conference*, Sept., 2006, pp. 311–314.

[101] K. Sanborn, D. Ma, and V. Ivanov, "A sub-1-V low-noise bandgap voltage reference," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 11, pp. 2466–2481, Nov., 2007.

[102] Y. Dai, D. T. Comer, D. J. Comer, and C. S. Petrie, "Threshold voltage based CMOS voltage reference," *IEE Proceedings - Circuits, Devices and Systems*, vol. 151, no. 1, pp. 58–62, Feb., 2004.

[103] Y. Yang, D. M. Binkley, L. Li, C. Gu, and C. Li, "All-CMOS subbandgap reference circuit operating at low supply voltage," in *Proceedings of the IEEE International Symposium of Circuits and Systems (ISCAS)*, May, 2011, pp. 893–896.

[104] B. A. I. U. K., S. Okura, T. Ido, and K. Taniguchi, "An area-efficient CMOS bandgap reference utilizing a switched-current technique," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 57, no. 10, pp. 762–766, Oct., 2010.

[105] Y. Okuda, T. Tsukamoto, M. Hiraki, M. Horiguchi, and T. Ito, "A trimming-free CMOS bandgap-reference circuit with sub-1-V-supply voltage operation," in *Proceedings of the IEEE Symposium on VLSI Circuits*, June, 2007, pp. 96–97.

[106] R. T. Perry, S. H. Lewis, A. P. Brokaw, and T. R. Viswanathan, "A 1.4V supply CMOS fractional bandgap reference," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2180–2186, Oct., 2007.

[107] J. Doyle, Y. J. Lee, Y.-B. Kim, H. Wilsch, and F. Lombardi, "A CMOS subbandgap reference circuit with 1-V power supply voltage," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 1, pp. 252–255, Jan., 2004.

[108] H. Zhang, X. Liu, J. Zhang, H. Zhang, J. Li, R. Zhang, S. Chen, and A. C. Carusone, "A nano-watt MOS-only voltage reference with high-slope ptat voltage generators," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. PP, no. 99, pp. 1–1, 2017.

[109] L. Magnelli, F. Crupi, P. Corsonello, C. Pace, and G. Iannaccone, "A 2.6 nW, 0.45V temperature-compensated subthreshold CMOS voltage reference," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 2, pp. 465–474, Feb., 2011.

[110] D. Albano, F. Crupi, F. Cucchi, and G. Iannaccone, "A sub-kT/q voltage reference operating at 150 mv," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 8, pp. 1547–1551, Aug, 2015.

[111] H. W. Huang, C. Y. Hsieh, K. H. Chen, and S. Y. Kuo, "A 1V 16.9ppm/°C 250nA switched-capacitor CMOS voltage reference," in *Proceedings of the IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, Feb., 2008, pp. 438–626.

[112] H. Wang and Q. Ye, "A CMOS voltage reference without resistors for ultra-low power applications," in *Proceedings of the International Conference on ASIC*, Oct., 2007, pp. 526–529.

[113] P. Kinget, C. Vezyrtzis, E. Chiang, B. Hung, and T. L. Li, "Voltage references for ultra-low supply voltages," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, Sept., 2008, pp. 715–720.

[114] Y. Osaki, T. Hirose, N. Kuroki, and M. Numa, "1.2-V supply, 100-nW, 1.09-V bandgap and 0.7-V supply, 52.5-nW, 0.55-V subbandgap reference circuits for nanowatt CMOS LSIs," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 6, pp. 1530–1538, June, 2013.

[115] E. M. C. Galeano, A. Olmos, and A. L. V. Boas, "A very low power area efficient CMOS only bandgap reference," in *2012 25th Symposium on Integrated Circuits and Systems Design (SBCCI)*, Aug., 2012, pp. 1–6.

[116] B. D. Yang, "250-mV supply subthreshold CMOS voltage reference using a low-voltage comparator and a charge-pump circuit," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 11, pp. 850–854, Nov., 2014.

[117] H. Wang and P. Mercier, "A 14.5pW, 31ppm/°C resistor-less 5 pA current reference employing a self-regulated push-pull voltage reference generator," in *Proc. of the IEEE International Symposium on Circuits and Systems*, May, 2016, pp. 1290–1293.

[118] T. Vergine, M. D. Matteis, S. Michelis, G. Traversi, F. D. Canio, and A. Baschirotto, "A 65nm rad-hard bandgap voltage reference for LHC environment," *IEEE Transactions on Nuclear Science*, vol. 63, no. 3, pp. 1762–1767, June, 2016.

[119] K. Ueno, T. Hirose, T. Asai, and Y. Amemiya, "A 300nW, 15$ppm/°C$, 20ppm/V CMOS voltage reference circuit consisting of subthreshold MOSFETs," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 2047–2054, July, 2009.

[120] P. H. Huang, H. Lin, and Y. T. Lin, "A simple subthreshold CMOS voltage reference circuit with channel- length modulation compensation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 9, pp. 882–885, Sept., 2006.

[121] Y. Ji, C. Jeon, H. Son, B. Kim, H. Park, and J. Sim, "A 9.3nW all-in-one bandgap voltage and current reference circuit," in *IEEE ISSCC Dig. Tech. Papers*, Feb., 2017, pp. 100–101.

[122] J. Lee, Y. Ji, S. Choi, Y. Cho, S. Jang, J. Choi, B. Kim, H. Park, and J. Sim, "A 29nW bandgap reference circuit," in *IEEE ISSCC Dig. Tech. Papers*, Feb., 2015, pp. 1–3.

[123] Q. Duan and J. Roh, "A 1.2-V 4.2- $ppm/°C$ high-order curvature-compensated CMOS bandgap reference," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 3, pp. 662–670, March, 2015.

[124] Z. K. Zhou, Y. Shi, Z. Huang, P. S. Zhu, Y. Q. Ma, Y. C. Wang, Z. Chen, X. Ming, and B. Zhang, "A 1.6-V 25-μA 5-$ppm/°C$ curvature-compensated bandgap reference," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 4, pp. 677–684, April, 2012.

[125] X. Ming, Y. q. Ma, Z. k. Zhou, and B. Zhang, "A high-precision compensated CMOS bandgap voltage reference without resistors," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 57, no. 10, pp. 767–771, Oct., 2010.

[126] M. D. Ker and J. S. Chen, "New curvature-compensation technique for CMOS bandgap reference with sub-1-V operation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 8, pp. 667–671, Aug., 2006.

[127] N. Alhasssan, Z. Zhou, and E. S. Sinencio, "An all-MOSFET sub-1-V voltage reference with a -51dB PSR up to 60 MHz," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 3, pp. 919–928, March, 2017.

[128] A. S. Cardoso, P. S. Chakraborty, N. Karaulac, D. M. Fleischhauer, N. E. Lourenco, Z. E. Fleetwood, A. P. Omprakash, T. D. England, S. Jung, L. Najafizadeh, N. J. H. Roche, A. Khachatrian, J. H. Warner, D. McMorrow, S. P. Buchner, E. X. Zhang, C. X. Zhang, M. W. McCurdy, R. A. Reed, D. M. Fleetwood, P. Paki-Amouzou, and J. D. Cressler, "Single-event transient and total dose response of precision voltage reference circuits designed in a 90-nm SiGe BiCMOS technology," *IEEE Transactions on Nuclear Science*, vol. 61, no. 6, pp. 3210–3217, Dec., 2014.

[129] A. J. Annema, P. Veldhorst, G. Doornbos, and B. Nauta, "A sub-1V bandgap voltage reference in 32nm FinFET technology," in *Proceedings of the IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, Feb., 2009, pp. 332–333.

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

[131] M. Al-Shyoukh and A. Kalnitsky, "A 500nA quiescent current, trim-free, ±1.75% absolute accuracy, CMOS-only voltage reference based on anti-doped n-channel MOSFETs," in *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*, Sept., 2014, pp. 1–4.

[132] K. Y. Wong, W. Chen, and K. J. Chen, "Integrated voltage reference generator for GaN smart power chip technology," *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 952–955, April, 2010.

[133] Y. Wang, Z. Zhu, J. Yao, and Y. Yang, "A 0.45-V, 14.6-nW CMOS subthreshold voltage reference with no resistors and no BJTs," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 7, pp. 621–625, July, 2015.

[134] L. Lu and C. Li, "Offset error reduction using gate-bulk-driven error correction amplifier for low-voltage sub-bandgap reference," *Electronics Letters*, vol. 49, no. 11, pp. 694–696, May, 2013.

[135] H. Zhuang, Z. Zhu, and Y. Yang, "A 19-nW 0.7-V CMOS voltage reference with no amplifiers and no clock circuits," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 11, pp. 830–834, Nov., 2014.

[136] J. H. Li, X. b. Zhang, and M. y. Yu, "A 1.2-V piecewise curvature-corrected bandgap reference in 0.5 μm CMOS process," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 6, pp. 1118–1122, June, 2011.

[137] P. Malcovati, F. Maloberti, M. Pruzzi, and C. Fiocchi, "Curvature compensated BiCMOS bandgap with 1V supply voltage," in *Proceedings of the 26th European Solid-State Circuits Conference*, Sept., 2000, pp. 7–10.

[138] P. B. Basyurt, E. Bonizzoni, D. Y. Aksin, and F. Maloberti, "A 0.4-V supply curvature-corrected reference generator with $84.5 ppm/^{\circ}C$ average temperature coefficient within -40°C to 130°C," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 4, pp. 362–366, April, 2017.

[139] H. M. Chen, C. C. Lee, S. H. Jheng, W. C. Chen, and B. Y. Lee, "A sub-1 ppm/°C precision bandgap reference with adjusted-temperature-curvature compensation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 6, pp. 1308–1317, June, 2017.

[140] Z. Zhu, J. Hu, and Y. Wang, "A 0.45V, Nano-Watt 0.033% line sensitivity MOSFET-only sub-threshold voltage reference with no amplifiers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 9, pp. 1370–1380, Sept., 2016.

[141] I. M. Filanovsky and A. Allam, "Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 7, pp. 876–884, July, 2001.

Appendix A

A 788mV voltage reference cell

A low-power voltage reference cell for system-on-chip applications is presented in this Appendix. A combination of thin-oxide transistors and thick-oxide transistors is used to generate a low temperature-coefficient (TC) reference voltage. We propose a new curvature compensation circuit to reduce the process variations of the output voltage. The circuit is fabricated in a standard 0.35μm CMOS process. The proposed circuit generates a $788V$ output and achieves a TC of 141ppm/°C from $-5°C$ to $110°C$. The line regulation is $0.9mV/V$ from 1.5V to 3.3V supply. The power consumption of this chip under a 3.3V supply is 0.4μW at room temperature. The die area of this chip is $0.399mm^2$ which includes the serial peripheral interface (SPI) circuits and trimming circuits.

A.1 The proposed circuit

Voltage reference circuits are critical building blocks that are used to generate a stable voltage across a wide range of temperatures. A low temperature-coefficient (TC), low power voltage reference circuit can potentially have many system-on-chip applications.

While good designs have been presented, these designs are not without issues. Some of the voltage reference cells are available in standard CMOS processes which consume high power (e.g. [85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 99, 105, 106, 107]). Line regulation is another important parameter. There are a few published low TC circuits which have poor line regulation [108, 109, 110, 111, 112, 113, 85, 111, 114, 115, 116, 110, 117, 102, 118]. The output voltage of a few voltage reference circuits is limited to the threshold voltage [119] of a transistor or is very small [120]. A number of other voltage references with good temperature performance and low power have been reported, but these circuits require special devices that may not be available in standard CMOS processes, such as NPN and PNP transistors [121, 122, 123, 124, 125, 126], $DTMOS$ transistors [127], $SG\ BiCMOS$ process [128], $SOI-FINFET$ [129], depletion-mode $NMOS$ transistors [130], anti-doped $NMOS$ devices [131], GaN process [132], anti-doped n-

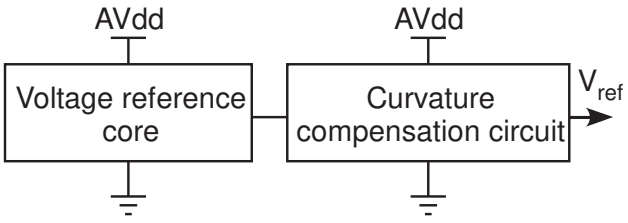


Figure A.1: Top level block diagram of the proposed voltage reference core

channel MOSFETs [131], and native nMOS devices [64]. [133], [65], and [134] are a few designs which use triple-well $CMOS$ processes. Other voltage references in standard CMOS processes with reduced power consumption may have temperature-coefficients that are too large for many applications [67, 68, 59]. Finally, some of the designs are extremely process-sensitive. For example, [135] has 100mV die-to-die variations without trimming.

In this Appendix, we focus on the design of a low-power, low-TC voltage reference cell with small dependency on the process variations in a 0.35μm standard CMOS technology. Many voltage reference circuits use curvature compensation circuits to generate a low-TC output voltage. [136, 137, 126, 138, 139] are a few examples which are not power efficient. Also, a few curvature compensation circuits are not designed in standard CMOS processes [140, 133, 126, 137, 138].

In this Appendix, we present a low-power curvature-compensation technique to reduce the TC of the voltage reference cell in a standard CMOS process. The process sensitivity is another important factor in the design of voltage reference cells. In this Appendix, we present a circuit that utilizes a process compensation circuit to reduce the sensitivity of the output voltage to process corners. Our voltage reference cell has been fabricated in a standard 0.35μm CMOS process and is capable of generating a 788mV reference voltage with a TC of <150ppm/°C and single μW power consumption over a wide range of temperatures. To achieve this performance, we use both standard (thin-oxide) transistors and I/O thick-oxide transistors to provide a low-TC reference voltage.

A.2 Principle of Operation

A high-level block diagram of the voltage reference cell is shown in Fig. A.1. In this circuit, a voltage reference core generates the reference DC voltage. Then a curvature compensation circuit removes the curvature from this voltage and a low TC output voltage is generated at the output of the circuit (V_{ref}). One advantage of this technique over other curvature compensation techniques is that this technique is applicable to other voltage reference cells, too.

To achieve low power consumption, our circuit was designed to operate in the subthreshold region with a very low bias currents. The drain current of a transistor biased in subthreshold can

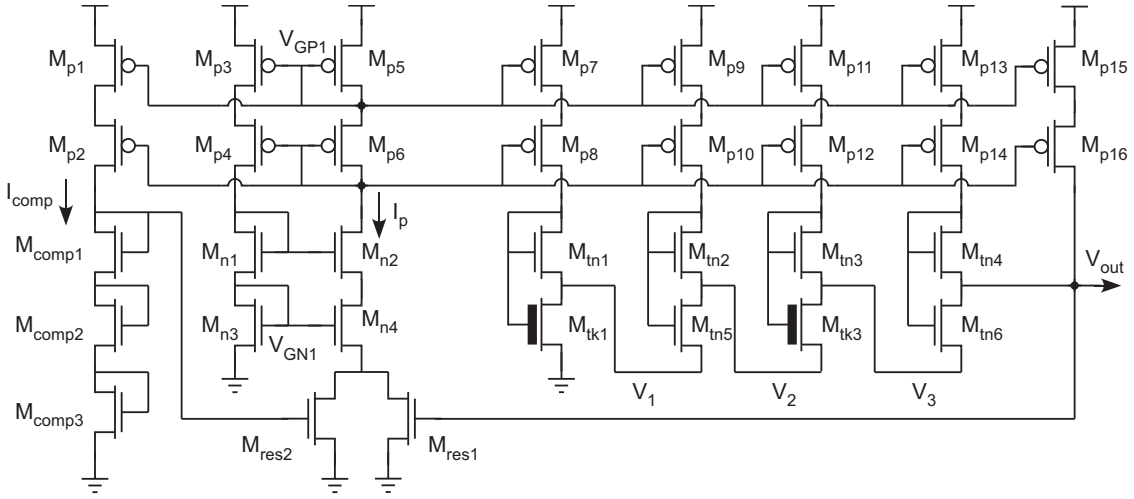


Figure A.2: Top level block diagram of the voltage reference cell

be expressed as [57]:

$$I_d = I_0' S e^{\frac{\kappa(V_g - V_T)}{U_T}} \left(e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}} \right) \tag{A.1}$$

where $I_0' = 2\eta\mu C_{ox} U_T^2$, S is the aspect ratio (W/L) of the transistor, $U_T = kT/q$ is the thermal voltage, k is the Boltzmann constant, q is the elementary charge, T is the absolute temperature in kelvins, V_T is the MOSFET threshold voltage, and $\eta = 1/\kappa$ is the subthreshold slope. Voltages V_g, V_s, and V_d are the gate, source, and drain voltages, respectively, referenced to the substrate. When the transistor operates in the saturation region, $exp(\frac{-V_d}{U_T})$ approaches zero and is, therefore, negligible. We also assume that the transistors have been designed with large-enough channel lengths that the channel-length modulation effect can be safely dropped.

A.2.1 Voltage reference core

The schematic diagram of the proposed voltage reference core is shown in Fig. A.2. This circuit consists of a current reference cell and a voltage generator. All transistors are biased in sub-threshold except for M_{res1} and M_{res2} which are biased in above threshold linear region. The gate-to-source voltage of M_{res1} and M_{res2} is high enough to bias these two in the above-threshold triode region. The current reference cell includes M_{p3}, M_{p4}, M_{p5}, M_{p6} as a one-to-one current mirror. This circuit also includes M_{n1}, M_{n2}, M_{n3}, and M_{n4} which are biased in the sub-threshold region. The V_{gs} difference between M_{n3} and M_{n4} is converted to a current using M_{res1} and M_{res2} acting as resistors.

The voltage generator circuit includes 4 stages. Two of these stages are CTAT sources and two of them are PTAT sources. The CTAT voltage generator is shown in Fig. A.3. This circuit has been used in [61]. M_{tk} is a thick-oxide 5V transistor and M_{tn} is a thin-oxide 3.3V transistor. The difference of the oxide thicknesses of these two transistors causes a threshold voltage difference.

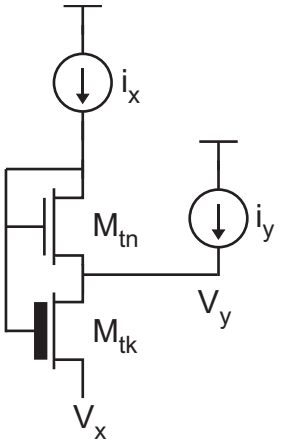


Figure A.3: Schematic diagram of the type 1 composite circuit

Therefore, the voltage difference produced by this cell between V_x and V_y can be expressed as:

$$V_{yx} = V_y - V_x = \kappa(V_{th,thick} - V_{th,thin}) + U_T ln\Big(\frac{S_{thin}}{S_{thick}}\frac{I_0}{I_{0t}}\frac{i_x + i_y}{i_x}\Big) \tag{A.2}$$

The threshold voltage difference of the transistors is a CTAT term. The ln term is a function of the transistor sizes and the bias current ratio. By proper sizing of the transistors, this block will become a CTAT voltage generator. κ has high-order temperature dependent terms and will cause up or down curvature in the output of the voltage generator based on the sign of the ln term.

The PTAT voltage generator is shown in Fig. A.4. This circuit has been used in a few voltage reference cells [127]. This circuit includes two thin-oxide transistors. The voltage difference produced by this cell between V_x and V_y can be expressed as:

$$V_{yx} = V_y - V_x = U_T ln\Big(\frac{S_{tn1}}{S_{tn2}}\frac{(i_x + i_y)}{i_x}\Big) \tag{A.3}$$

By choosing proper transistor sizes, the term inside the ln can be made much greater than 1 and hence a PTAT voltage will be generated.

The voltage at the intermediate nodes of the voltage generator of Fig. A.2 can be calculated as follows. The size of all the mirror transistors (M_{p7} to M_{p16}) are the same. The bias currents for the first CTAT stage are i_x and $i_y = 4i_x$ and the voltage drop across this stage can be expressed as:

$$V_1 = \kappa(V_{th,tk1} - V_{th,tn1}) + U_T ln\Big(\frac{5S_{tn1}}{S_{tk1}}\frac{I_0}{I_{0t}}\Big) \tag{A.4}$$

The bias currents for the first PTAT stage are i_x and $i_y = 3i_x$ and the voltage drop across this

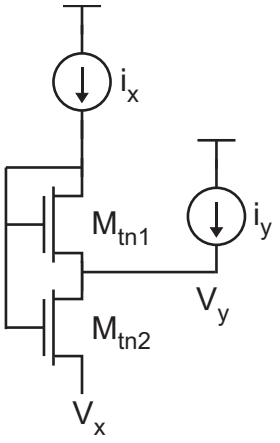


Figure A.4: Schematic diagram of the type 2 composite circuit

stage can be expressed as (V_{21}):

$$V_{21} = V_2 - V_1 = U_T ln\Big(\frac{4S_{tn2}}{S_{tn5}}\Big) \tag{A.5}$$

The bias currents for the second CTAT stage are i_x and $i_y = 2i_x$ and the voltage drop across this stage can be expressed as (V_{32}):

$$V_{32} = V_3 - V_2 = \kappa(V_{th,tk3} - V_{th,tn3}) + U_T ln\Big(\frac{3S_{tn3}}{S_{tk3}\frac{I_0}{I_{0t}}}\Big) \tag{A.6}$$

Finally, the bias currents for the second PTAT stage are i_x and $i_y = i_x$ and the voltage drop across this stage can be expressed as ($V_{out} - V_3$):

$$V_{out} - V_3 = U_T ln\Big(\frac{2S_{tn4}}{S_{tn6}}\Big) \tag{A.7}$$

Adding all of these equations (A.4, A.5, A.6, and A.7), and assuming almost equal threshold voltages for the transistors, the output voltage can be expressed as:

$$V_{out} = 2\kappa(V_{th,tk} - V_{th,tn}) + U_T ln(\frac{120S_{tn1}S_{tn2}S_{tn3}S_{tn4}}{S_{tk1}S_{tk3}S_{tn5}S_{tn6}}\frac{I_0^2}{I_{0t}^2}) \tag{A.8}$$

The output voltage V_{out} with a zero TC can be obtained by adjusting the size of the transistors. The CTAT part of this voltage is $\kappa(V_{th,tk} - V_{th,tn})$ and the PTAT part of this volatge is $U_T ln(\frac{120S_{tn1}S_{tn2}S_{tn3}S_{tn4}}{S_{tk1}S_{tk3}S_{tn5}S_{tn6}}\frac{I_0^2}{I_{0t}^2})$. The temperature dependence of the threshold voltage can be given by

$$V_{th} = V_{th0} - \alpha T \tag{A.9}$$

where V_{th0} is the threshold voltage at 0K, and κ is the TC of V_{th} [141]. Sign of the ln() term can be set by using proper values for the transistor.

Again, κ will produce a high order temperature term which creates the up or down curvature. Using equation A.9, A.8 could be rewritten as:

$$V_{out} = 2\kappa(V_{th0,tk} - \alpha_{tk}T - V_{th0,tn1} + \alpha_{tn}T) + U_T ln(\frac{120 S_{tn1}S_{tn2}S_{tn3}S_{tn4}}{S_{tk1}S_{tk3}S_{tn5}S_{tn6}} \frac{I_0^2}{I_{0t}^2}) \tag{A.10}$$

The threshold voltage of a transistor is a CTAT term. The slope of the threshold voltage against temperature is higher for thick oxide transistors. Thus the first term ($2\kappa(V_{th0,tk} - \alpha_{tk}T - V_{th0,tn1} + \alpha_{tn}T)$) in (A.10) is a CTAT voltage. The ln term in (A.10) can be set by the aspect ratios of the transistors and can be a high slope PTAT voltage. Thus, this circuit can generate either a CTAT or a PTAT output voltage depending on the aspect ratios of the transistors and the bias currents going through them.

A.2.2 Process compensation circuit

The proposed current reference cell has a similar structure to the conventional current reference cells except for the resistive element. A transistor biased in the above threshold linear region is used in [114, 133] instead of a resistor to generate the bias current. In this work, we follow the same idea, but we try to reduce the process sensitivity of the current reference, too.

Fortunately, unlike [119] both V_{out} and TC_{Vout} are not functions of the bias current. However, the region of operation for the transistors in the voltage reference generator (M_{tn1} M_{tn6}, M_{tk1}, and M_{tk3}) are a function of their bias currents, and we must make sure that all these transistors are biased in subthreshold region. Otherwise, we can not use the analysis presented in the previous section to find V_{out}. Therefore, we must reduce the process sensitivity of the bias current. The bias currents of these cells are a function of I_p. Accordingly, we must make sure that this current stays in an acceptable range. I_p is a function of the voltage difference between M_{n3} and M_{n4} and the resistance seen across M_{res1}. Based on Cadence simulations, a big portion of variations in I_p across corners is due to the process variations of M_{res1}. In this section we propose a technique to reduce the process variations of I_p. The voltage reference core without process compensation circuit is shown in (A.5). The currents being conducted through two branches of the current reference cell are equal, because:

$$(w/l)_{p3} = (w/l)_{p4} = (w/l)_{p5} = (w/l)_{p6} \tag{A.11}$$

A KVL equation around V_{GN1} gives this:

$$V_{GS,n3} = V_{GS,n4} + V_{DS,res1} \tag{A.12}$$

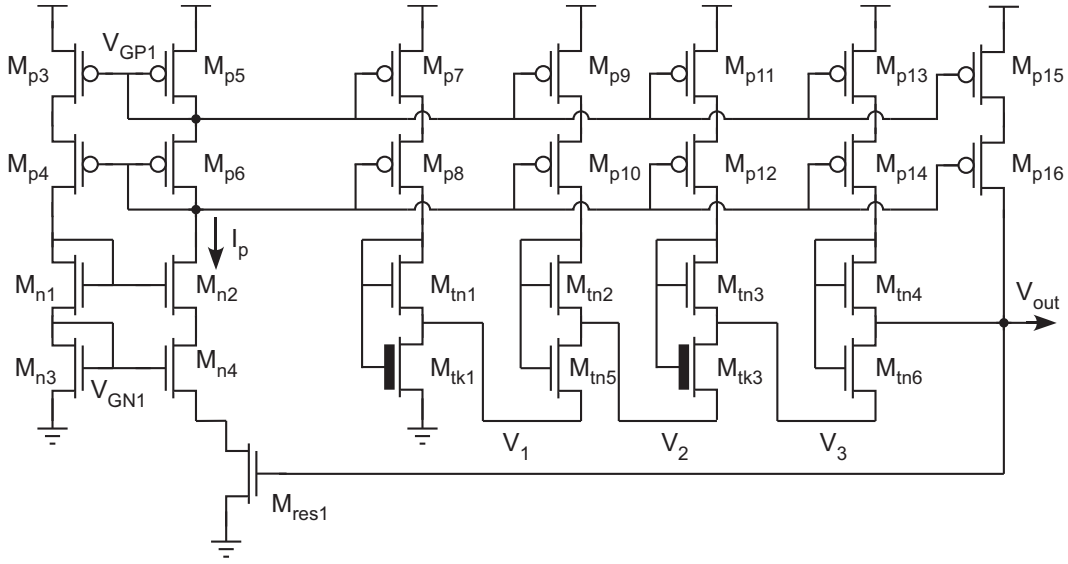


Figure A.5: Schematic diagram of the voltage reference core using conventional MOS resistor.

Since, M_{n3} and M_{n4} have the same current, we will have:

$$V_{DS,res1} = \eta U_T ln(\frac{(w/l)_{n4}}{(w/l)_{n3}}) \tag{A.13}$$

MOS resistor M_{res1} is biased in strong inversion, deep triode region. Therefore, its resistance is given by:

$$R_1 = \frac{1}{2\mu_n C_{ox}(w/l)_{res1}(V_{out} - V_{th,res1})} \tag{A.14}$$

Accordingly, the current generated by the current reference cell could be expressed as:

$$I_p = \frac{V_{DS,res1}}{R_1} = 2\mu_n C_{ox}(w/l)_{res1}(V_{out} - V_{th,res1})\eta U_T ln(\frac{(w/l)_{n4}}{(w/l)_{n3}}) \tag{A.15}$$

Equation (A.15) clearly shows that I_p is process dependent. In this equation μ, C_{ox}, V_{th}, V_{out} are process-dependent. We would like to suppress the process dependency of this current. The process dependence of I_p on the threshold voltage is much higher than other parameters [119]. The compensation circuit includes M_{p1}, M_{p2}, M_{comp1}, M_{comp2}, M_{comp3}, and M_{res2}. The bias current I_{comp} was chosen to be much smaller than I_p. Therefore, it will not affect the total power consumption of the chip. The aspect ratio of M_{comp1}, M_{comp2}, and M_{comp3} is selected such that M_{res2} is biased in above threshold triode region. Therefore, its resistance is given by:

$$R_2 = \frac{1}{2\mu_n C_{ox}(w/l)_{res2}(V_{GS,res2} - V_{th,res2})} \tag{A.16}$$

The gate-source voltage of M_{res2} can be expressed as:

$$V_{GS,res2} = V_{GS,comp1} + V_{GS,comp2} + V_{GS,comp3} \tag{A.17}$$

Therefore,

$$V_{GS,res2} = V_{th,comp3} + \eta U_T ln(\frac{I_{comp}}{S_{comp3}\mu U_T^2}) + V_{th,comp2} + \eta U_T ln(\frac{I_{comp}}{S_{comp2}\mu U_T^2}) + V_{th,comp1} + \eta V_T ln(\frac{I_{comp}}{S_{comp1}\mu V_T^2}) \tag{A.18}$$

This equation could be simplified to the following:

$$V_{GS,res2} = V_{th,comp3} + V_{th,comp2} + V_{th,comp1} + \eta U_T ln(\frac{I_{comp}^3}{\mu^3 U_T^6 S_{comp1} S_{comp2} S_{comp3}}) \tag{A.19}$$

Therefore, the resistance of M_{res2} can be expressed as:

$$R_2 = \frac{1}{2\mu_n C_{ox}(w/l)_{res2}(V_{th,comp3} + V_{th,comp2} + V_{th,comp1} + \eta U_T ln(\frac{I_{comp}^3}{\mu^3 U_T^6 S_{comp1} S_{comp2} S_{comp3}}) - V_{th,res2})} \tag{A.20}$$

The parallel combination of R_1 and R_2 can be expressed as:

$$R_{eq} = \frac{1}{\frac{1}{R1} + \frac{1}{R2}} \tag{A.21}$$

This equivalent resistance can be expressed as:

$$R_{eq} = \frac{1}{2\mu_n C_{ox}((w/l)_{res1}(V_{out} - V_{th,res1}) + (w/l)_{res2}(\Sigma V_{th,comp} + \eta U_T ln(\frac{I_{comp}^3}{\mu^3 U_T^6 S_{comp1} S_{comp2} S_{comp3}}) - V_{th,res2})} \tag{A.22}$$

And can be simplified to the following:

$$R_{eq} = \frac{1}{2\mu_n C_{ox}(w/l)_{res1}(V_{out} - V_{th,res1} + \frac{(w/l)_{res2}}{(w/l)_{res1}}(\Sigma V_{th,comp} + \eta U_T ln(\frac{I_{comp}^3}{\mu^3 U_T^6 S_{comp1} S_{comp2} S_{comp3}}) - V_{th,res2})} \tag{A.23}$$

The denominator of (A.14) is a function of $V_{out} - V_{th,res1}$, which is process-sensitive. Process variations of $V_{out} - V_{th,res1}$ change the absolute value of the resistance. Hence, the current generated by

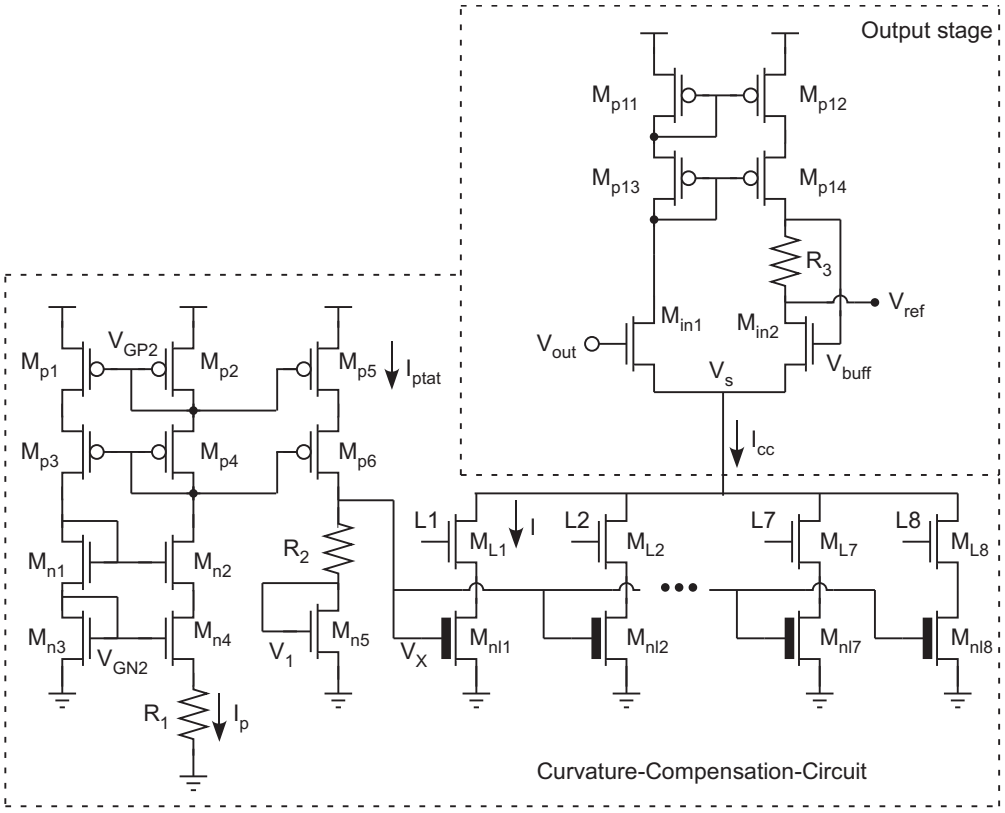


Figure A.6: Schematic diagram of the curvature compensation circuit.

this cell varies among corners. However, in (A.20), the process variations can be compensated by the second term which is added to the denominator ($\frac{(w/l)_{res2}}{(w/l)_{res1}}(\Sigma V_{th,comp}+\eta U_T ln(\frac{I_{comp}^3}{\mu^3 U_T^6 S_{comp1} S_{comp2} S_{comp3}})-V_{th,res2}$). There are enough parameters in this equation, which gives a high level of freedom to achieve V_{th} process compensation ($(w/l)_{res1}, S_{comp1}, S_{comp2}, S_{comp3}, I_{comp}$). There are some other process-dependent parameters (μ_n, C_{ox}) in this equation, too. However, we just focus on V_{th} process variations, because compared to the other parameters mentioned in the past few lines, it has the highest rate of variations among corners.

A.3 Curvature compensation idea

Many curvature compensation circuits have been proposed in the literature. In this work, a flexible curvature compensation technique is presented that can be applied to other voltage reference circuits with high-temperature variations.

A.3.1 Curvature compensation circuit

The circuit diagram of the curvature compensation circuit is shown in Fig. A.6. This circuit has two main parts. The curvature compensation circuit that generates I_{cc} and the output stage which receives V_{out} from A.2 and generates V_{ref} (curvature compensated output).

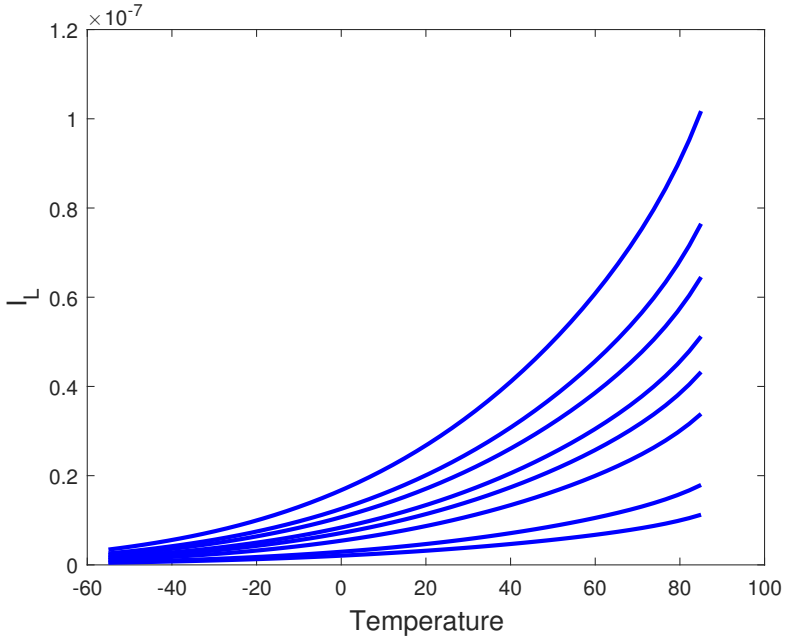


Figure A.7: Simulation results for I_L current while only one of the switches (L1-L8) is on at a time.

A.3.2 High-temperature non-linear current generation circuit

This circuit generates a nonlinear PTAT current that is close to zero at very low temperatures and rises to higher currents exponentially in higher temperatures. The current reference cell is a conventional current reference cell including M_{p1}, M_{p2}, M_{p3}, M_{p4}, M_{n1}, M_{n2}, M_{n3}, M_{n4}, and R_1. The current generated by this circuit is a PTAT current (I_P). This current is mirrored through M_{p5} and M_{p6} and generates another PTAT current called I_{ptat}. Transistors M_{nl1} to M_{nl8} are all thick-oxide transistors to make sure that they will be biased in sub threshold region. All other transistors are thin-oxide transistors. The gate voltage of the thick-oxide transistors can be expressed as:

$$V_x = R_2 I_{ptat} + V_1 \tag{A.24}$$

V_1, the gate voltage of M_{n5}, can be expressed as:

$$V_1 = V_{th,n5} + \frac{U_T}{\kappa} ln(\frac{I_{ptat}}{S_{n5} I_0}) \tag{A.25}$$

Therefore, I (current through M_{L1}) can be expressed as:

$$I = I_{0t} S_{nl1} exp\left(\frac{\kappa(V_x - V_{th,thick})}{U_T}\right) = S_{nl1} I_{0t} exp\left(\frac{\kappa(R_2 I_{ptat} + V_1 - V_{th,thick})}{U_T}\right) \tag{A.26}$$

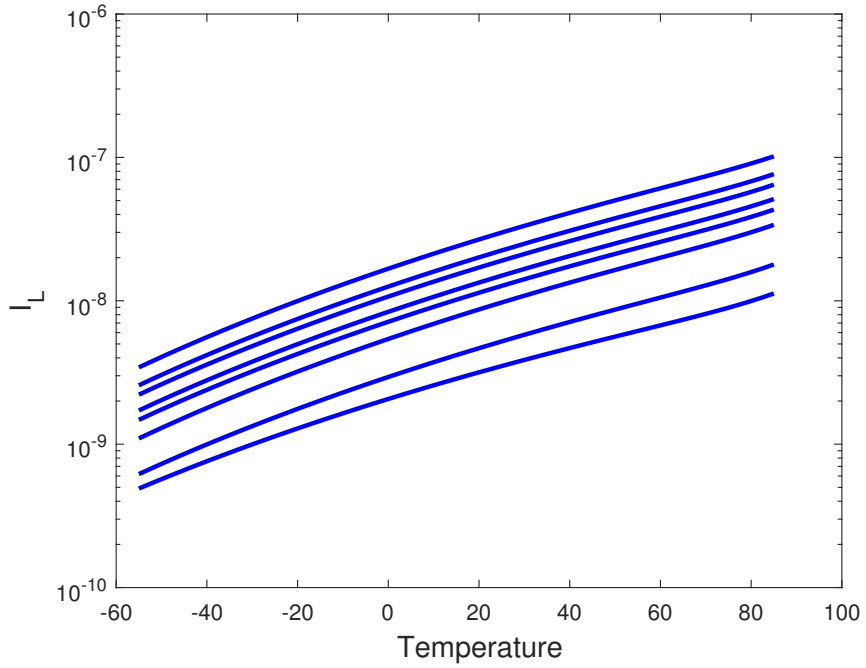


Figure A.8: Simulation results of $log(I_L)$ while only one of the switches (L1-L8) is on at a time.

This equation can be further simplified to the following:

$$I = S_{nl1} I_{0t} exp\left(\frac{\kappa(R_2 I_{ptat} + V_{th,n5} + \frac{U_T}{\kappa} ln(\frac{I_{ptat}}{S_{n5} I_0}) - V_{th,thick})}{U_T}\right) \tag{A.27}$$

Therefore, I can be further simplified to the following:

$$I = S_{nl1} I_{0t} exp\left(\frac{\kappa(V_{th,n5} - V_{th,thick})}{U_T}\right) exp\left(\frac{\kappa R_2 I_{ptat}}{U_T}\right) exp\left(ln\left(\frac{I_{ptat}}{S_{n5} I_0}\right)\right) \tag{A.28}$$

Again, this equation can be further simplified to this:

$$I = (S_{nl1} I_{0t})\left(\frac{I_{ptat}}{S_{n5} I_0}\right) exp\left(\frac{\kappa(V_{th,n5} - V_{th,thick})}{U_T}\right) exp\left(\frac{\kappa R_2 I_{ptat}}{U_T}\right) \tag{A.29}$$

$$I = I_{ptat} \frac{I_{0t}}{I_0} \frac{S_{nl1}}{S_{n5}} exp\left(\frac{\kappa(V_{th,n5} - V_{th,thick})}{U_T}\right) exp\left(\frac{\kappa R_2 I_{ptat}}{U_T}\right) \tag{A.30}$$

This current is very small for low temperatures and rises very fast with temperature. The $exp(\frac{\kappa(V_{th,n5} - V_{th,thick})}{U_T})$ term is almost independent of temperature since both the nominator and denominator are first-order functions of temperature. The rest of (A.30) is temperature-dependent. In $exp(\frac{RI_{ptat}}{\eta U_T})$, the RI_{ptat} term has high-order dependence on temperature, and U_T is first-order temperature dependent. Therefore, $I_{ptat} exp(\frac{\kappa R_2 I_{ptat}}{U_T})$ grows very fast with temperature. M_{L1}-M_{L8} are digital switches to control the amplitude of I_L. M_{nl1} to M_{nl8} are all thick-oxide transistors with different aspect

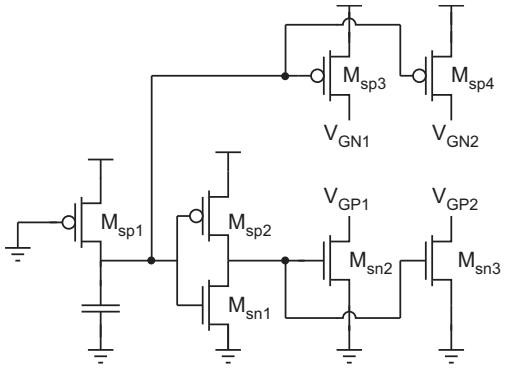


Figure A.9: Schematic diagram of the start-up circuit.

ratios. This will scale the current through each one of M_{L1}-M_{L8} up or down (look at equation (A.30)).

A.3.3 Output stage

The current through M_{in1} and M_{in2} can be expressed as:

$$I_{in1} = I_0 S_{in1} exp\left(\frac{\kappa(V_{out} - V_T)}{U_T}\right) exp\left(\frac{-V_S}{U_T}\right) \tag{A.31}$$

$$I_{in1} = I_0 S_{in1} exp\left(\frac{\kappa(V_{buff} - V_T)}{U_T}\right) exp\left(\frac{-V_S}{U_T}\right) \tag{A.32}$$

The aspect ratios of M_{p11}, M_{p12}, M_{p13}, and M_{p14} are all equal. Hence, the current through M_{in1} and M_{in2} are equal to $I_{cc}/2$. Therefore,

$$V_{buff} = V_{out} \tag{A.33}$$

Therefore, V_{buff} is an exact copy of V_{out}. The output voltage of this stage can be expressed as:

$$V_{ref} = V_{buff} - \frac{R_3 I_{cc}}{2} = V_{out} - \frac{R_3 I_{cc}}{2} \tag{A.34}$$

In order to compensate the curvature in V_{out}, the high-order terms in $R_3 I_{cc}/2$ must cancel the high-order terms in V_{out}.

A.3.4 Start-up circuit

The schematic diagram of the start-up circuit is shown in Fig. A.9. When the supply voltage is turned on, M_{sp1} starts to charge the capacitor slowly because the gate terminal is at ground. Therefore, the gate terminals of M_{sp3}, M_{sp4}, and M_{sp5} start to charge-up slowly. Accordingly, V_{GN1}, V_{GN2}, and V_{GN3} start to drop from V_{dd} to ground and will inject a current onto the shared

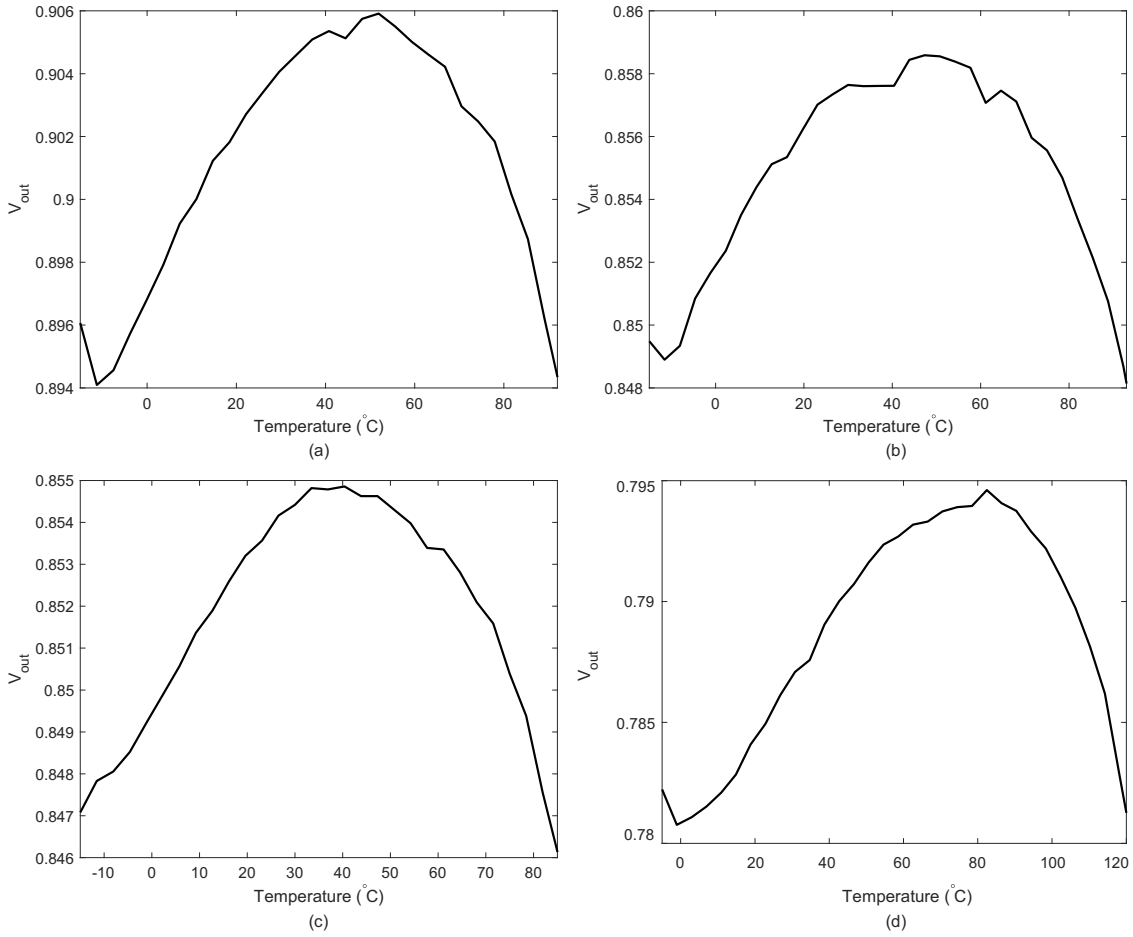


Figure A.10: Trimmed output for (a) TT=1 FF=SS=0 (125 ppm/°C from -11°C to 92°C) (b) TT=1 FF=SS=0 (94 ppm/°C from -15°C to 82°C) (c) TT=FF=SS=0 (125 ppm/°C from -15°C to 92°C) (d) TT=0 FF=SS=1 (139 ppm/°C from -5°C to 120°C).

gate connections of M_{n3}, M_{n4} (V_{GN1}) in the voltage reference core, the shared gate connection of M_{n3}, M_{n4} (V_{GN2}) in the curvature-compensation circuit, and the shared gate connection of M_{r1}, M_{r2} (V_{GN3}) in curvature compensation circuit. At the same time, the shared gate connections of M_{sn2}, M_{sn3}, and M_{sn4} start to drop from V_{dd} to ground. Hence, V_{GP1}, V_{GP2}, and V_{GP3} start to rise and M_{sn2}, M_{sn3}, and M_{sn4} inject some current to the gate connections of PMOS transistors in the current reference cells. Accordingly, the bias currents will not stabilize in a zero bias state.

A.4 Experimental results

This circuit was fabricated in a standard 0.35μm CMOS process. The circuit uses a shift register to feed in the digital data serially. This serial data includes $L_1 - L_8$, TT, FF, and SS. Widths of M_{p7}-M_{p16} are controllable with three trimming bits (TT, FF, SS). Basically each one of these transistors are a parallel combination of three separate transistors which their gates are controlled with TT, SS, and FF. Setting each one of these bits to zero will increase the W/L ratio of all M_{p7}-M_{p16}.

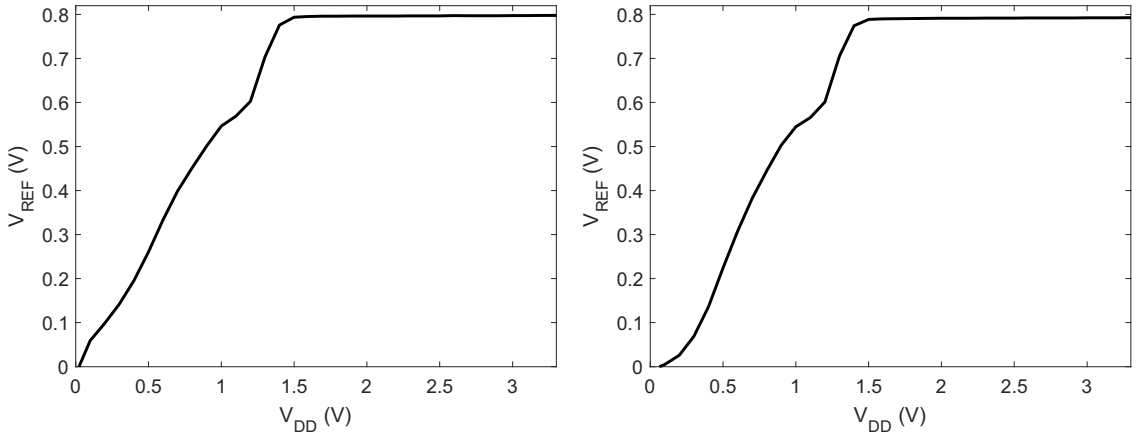


Figure A.11: Measured line regulation at room temperature for (a) Fail-Safe case LR=1.3mV/V (b) Trimmed case LR=0.9mV/V.

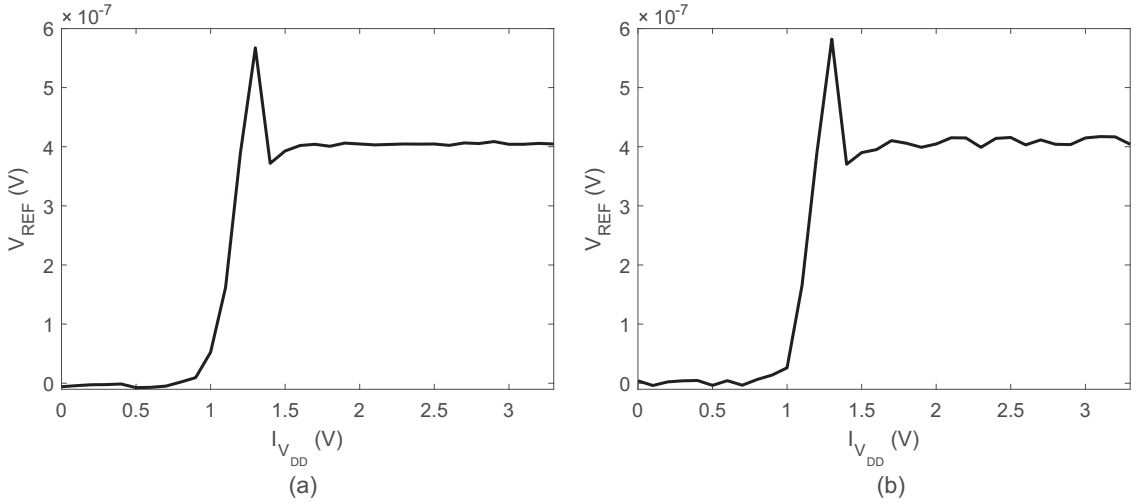


Figure A.12: (a) Measured current vs. supply voltage at room temperature. (b) Measured current vs. supply voltage at room temperature (Trimmed).

By increasing I_{cc}, the high-order curvature of the output voltage reduces and the operational temperature range of the circuit reduces too. Figure A.10 shows the output voltage for different trimming codes. The lowest TC is achievable when TT=1 and SS=FF=0. The temperature coefficient under this set-up is 94 ppm/°C. However, the operational temperature range of the circuit is -15°C to 82°C. The widest temperature range is achievable when TT=0 and FF=SS=1. However, the temperature coefficient of the circuit is 139 ppm/°C from -5°C to 120°C.

The line regulation measurements results of the proposed circuit before trimming and after trimming is shown in Fig. A.11 (a) and A.11(b), respectively. The supply voltage range of the proposed circuit is from $1.5V$ to $3.3V$. The measured supply current across the supply voltage is shown in Fig. A.12(a) and A.12(b), respectively. The current consumed by this chip in steady state for both cases are around $420nA$.

The voltage distribution of the proposed circuit for 10 different samples is shown in Fig. A.13. The circuit was set to TT=0 and FF=SS=1 and the DC output voltage of the circuit was measured

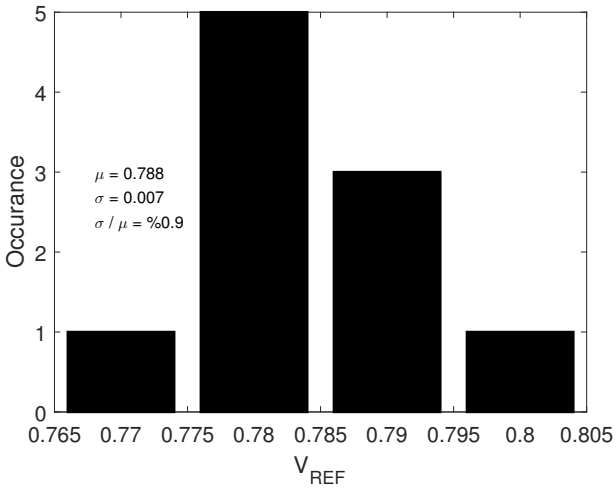


Figure A.13: Distribution of the DC output voltage for 10 different samples.

for these 10 samples. The average output voltage for these 10 different samples are $0.788mV$ and the σ/μ for these 10 samples is around %0.9.

A.4.1 Down curvature compensation

The curvature compensation idea could be generalized to cancel the down curvature, too. The same curvature compensation circuit is slightly changed to cancel the down curvature. In this case, instead of subtracting a voltage from V_{out} to generate V_{ref}, we add a voltage to V_{out} and generate V_{ref} to cancel down curvature. This circuit is shown in figure A.14. V_{buff} is an exact copy of V_{out} at all temperatures. A voltage drop across R_3 with up curvature is added to V_{buff} to cancel the down curvature visible in V_{buff}. The curvature compensated voltage is called V_{ref} in this circuit.

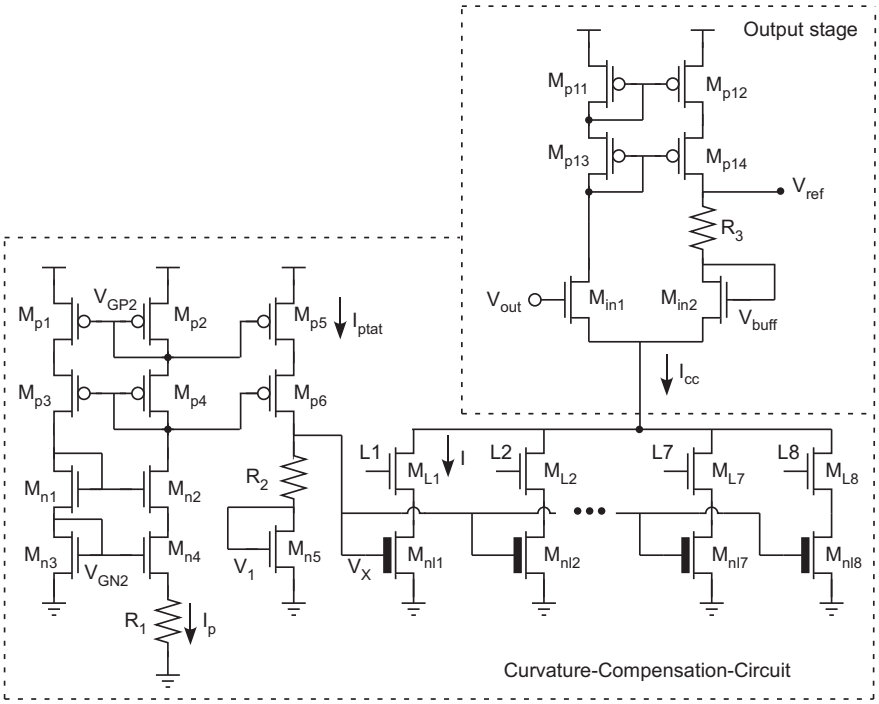


Figure A.14: The proposed idea to cancel down curvature.

Appendix B

A simple low power voltage reference cell

In chapter 4, we proposed a low-power voltage reference cell. We described the analysis of the circuit in broad terms. In this Appendix, we present a simple voltage reference cell which is even simpler than the voltage reference cell presented in chapter 4 but will provide a lower reference voltage.

B.1 The voltage reference cell

The schematic diagram of the circuit is presented in Fig. B.1. The operation of the start-up circuit and the PTAT current generator were explained in chapter 4. The output stage of this circuit includes a resistor (R_{out}) and a diode-connected transistor (M_Z). The voltage generated across the resistor is a PTAT voltage. The current (I_y) is a PTAT current, and it can be expressed as:

$$I_y = U_T \left(\frac{1}{R_C}\right) \left(\frac{S_{13}}{S_7}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{B.1}$$

Thus, the voltage across the resistor can be expressed as follows:

$$V_{PTAT} = U_T \left(\frac{R_{out}}{R_C}\right) \left(\frac{S_{13}}{S_7}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{B.2}$$

Using (4.1) and assuming that the transistor is biased in the subthreshold region, the voltage generated by the diode-connected transistor is given by:

$$V_x = V_{th} + \frac{U_T}{\kappa} \ln\left(\frac{I_d}{S_Z I_0'}\right) \tag{B.3}$$

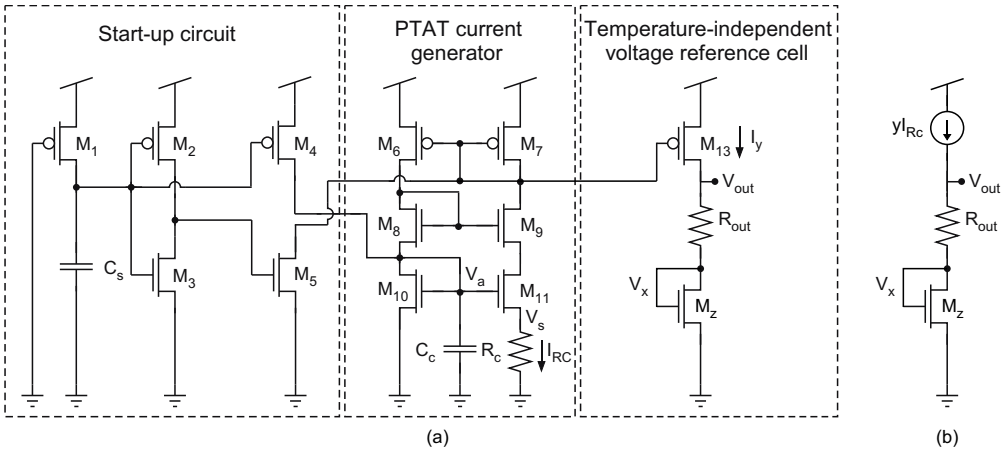


Figure B.1: Top level block diagram of the voltage reference cell

The threshold voltage term in this equation is a CTAT term. The second term is a function of U_T which is a PTAT term. This term can be set to be small by choosing a large aspect ratio (S_Z). Thus, the output voltage can be expressed as:

$$V_{out} = V_{th} + \frac{U_T}{\kappa} \ln\left(\frac{I_d}{S_Z I'_0}\right) + U_T \left(\frac{R_{out}}{R_C}\right)\left(\frac{S_{13}}{S_7}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{B.4}$$

The threshold voltage of a transistor is a CTAT term and can be expressed as:

$$V_{th} = \alpha_{V_{th}} T + V_{th0} \tag{B.5}$$

where $\alpha_{V_{th}}$ is a negative term. By taking the derivative of (B.4) with respect to T, we arrive at an equation for the TC at

$$TC = \frac{\partial V_{out}}{\partial T} = \alpha_{V_{th}} + \frac{K}{q\kappa} \ln\left(\frac{I_d}{S_Z I'_0}\right) + \frac{K}{q}\left(\frac{R_{out}}{R_C}\right)\left(\frac{S_{13}}{S_7}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{B.6}$$

The temperature coefficient of the output voltage can be set to zero by proper selection of S_Z, S_7, S_{13}, S_{10}, S_{11}, R_{out}, and R_C. Thus $\alpha_{V_{th}}$ can be expressed as:

$$\alpha_{V_{th}} = -\frac{K}{q\kappa} \ln\left(\frac{I_d}{S_Z I'_0}\right) - \frac{K}{q}\left(\frac{R_{out}}{R_C}\right)\left(\frac{S_{13}}{S_7}\right) \ln\left(\frac{S_{11}}{S_{10}}\right) \tag{B.7}$$

By using (B.7) in (B.4), the output voltage of this circuit can be simplified to the following:

$$V_{out} = V_{th0} \tag{B.8}$$

Therefore this circuit generates a voltage equal to the threshold voltage of a transistor at the reference temperature.