TEMPERATURE ROBUST PROGRAMMABLE SUBTHRESHOLD CIRCUITS THROUGH A BALANCED FORCE APPROACH

A Thesis
Presented to
The Academic Faculty

by

Brian P. Degnan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2013

TEMPERATURE ROBUST PROGRAMMABLE SUBTHRESHOLD CIRCUITS THROUGH A BALANCED FORCE APPROACH

Approved by:

Professor Jennifer Hasler,
Committee Chair
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Brad Minch
Department of Engineering
Olin

Professor Jennifer Hasler, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor David Anderson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Maysam Ghovanloo
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 4 APR 2011

ACKNOWLEDGEMENTS

I would like to thank the wonderful faculty at Georgia Tech, who made my time here so interesting, whether I wanted it to be or not. I would like to thank the members of both Dr. Jen Hasler's and Dr. David Anderson's research groups for being a great bunch. I particularly would like to thank the students that I've had, for keeping me on my toes. Finally, I would like to thank those with whom I may think differently, particularly **鄭路**, Richie Wunderlich, **田中大介**,**渡邊智之**, Aaron Nelson, and **前田水銀**.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The region of weak inversion for MOSFET operation is explored across temperature, application and function. A model has been developed that describes the first-order behavior of MOSFETs across all regions of operation and has been used to describe digital electrical behavior, estimate energy consumption, and as the basis for temperature-robust circuit design. These applications include programmable-transistor approaches for digital inverters, and programmable switch behavior for cross-bar switch matrices. Furthermore, the operational bounds of strong inversion operation are predicted. Finally, the method for describing transistor operation based upon force balance is presented.

CHAPTER I

INTRODUCTION

The MOSFET has several regions of operation with above-threshold operation being of primary concern. The region below threshold has generally just been a footnote; however, subthreshold operation has several desirable characteristics, such as simple physics, high-transconductance, and "sharp-edge" transitions in digital applications. In contrast, the temperature dependence of diffusion movement due to the thermal voltage terms is substantially greater than with super-threshold designs that operate through drift movement. This work aims to show architectures that are robust to temperature variations through symmetry, as well as integrate floating-gate transistors for precise programming of currents.

1.1 Subthreshold rises to the top

The digital designer is driven by the incentive of "faster" designs. The megahertz race of the nineties and naughties, and the constant need for speed driven by the desire for faster software, has resulted in designers rewarding complexity over simplicity. Transistor scaling has done an excellent job of giving similar, but not perfectly scaled, transistor changes with feature size. Regarding scaling, one often hears references to "Moore's Law", but the concept of scaling is generally not understood by those who casually throw around the term [43]. The idea of scaling and "Moore's Law" is simply the fact that more components can be squeezed onto an area of silicon. Previous work has demonstrated that scaling has not improved the performance of DSP regarding power with a million multiply accumulate (MMAC) operations stuck at about 4 milliwatts [39]. The processors are indeed "faster"; however, the performance increase is proportional to power consumption. ITRS 2009 roadmap also predicts this trend:

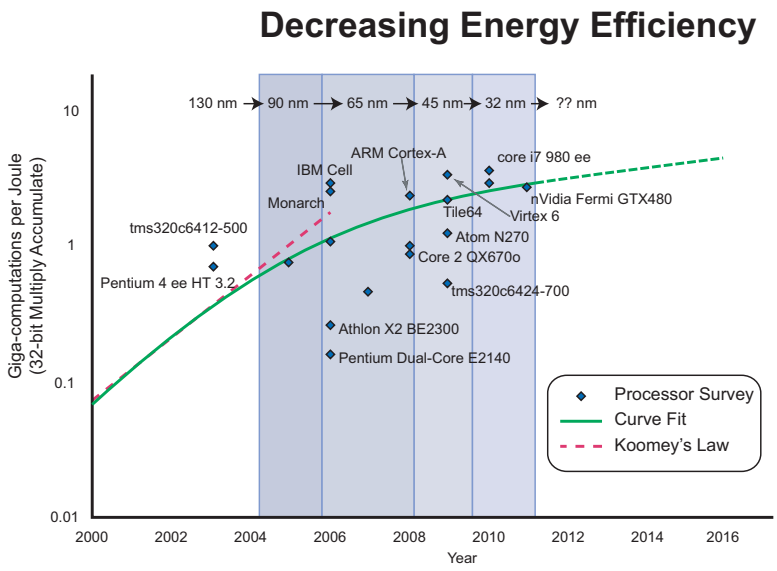


Figure 1: The transistor has not held its processing capability with the scaling of feature size. Using the power required for a multiply-accumulate function as a figure of merit for computation, modern processors are steadily decreasing in computational efficiency. The question is the trend on this efficiency, and whether it will be asymptotic.

'The subthreshold source/drain leakage current, I_{sd}, leak, is fixed at a value of 100 nA/μm for all years, which has important consequences for the chip power dissipation" [27]. A survey of available processors also shows a saturation in computational efficiency. Figure 1 illustrates this trend with a "computation" defined as a 32-bit multiply accumulate (MAC) operation [24, 1, 26, 46, 32]. As the graph shows, the data suggests that scaling is steadily decreasing computational efficiency of digital processors. The larger question is whether this decreasing computational efficiency will trend toward an efficiency barrier asymptotically, or just have a decreased slope when compared to previous generations.

In order to leverage all transistor operational regions, a unified model for power that integrates the MOSFET operational regimes must be created, which can be used to identify the corner-case behavior of short-circuit and leakage power estimates. The motivation for this is two-fold: a unified energy model is needed, and then power

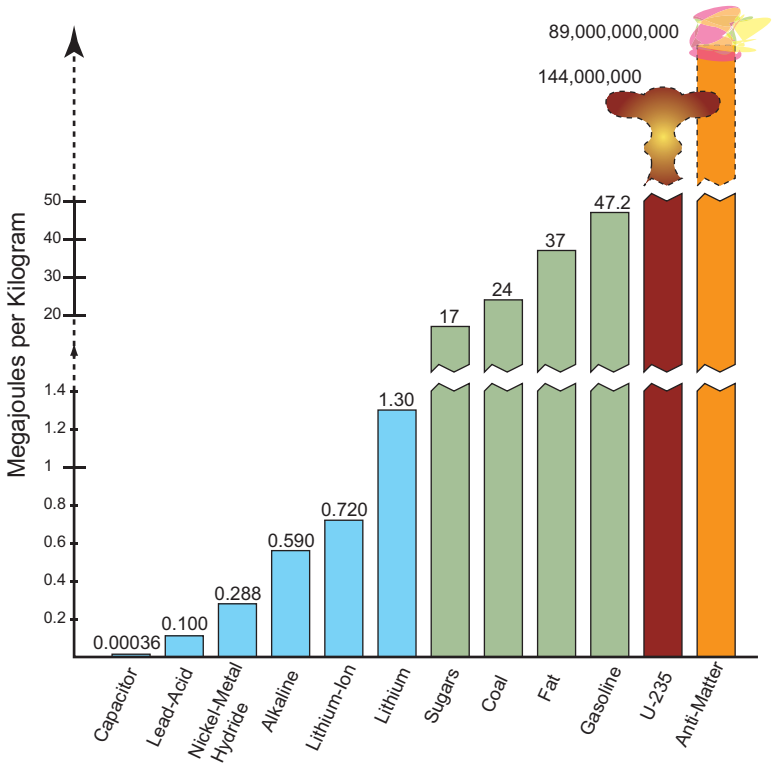


Figure 2: Energy density governs power constrained computing. The graph above shows theoretical energy density in megajoules per kilogram without considering loses due to conversion. Energy density of batteries has not increased as devices scaled to smaller feature sizes with higher leakage currents.

estimates can be made regardless of operating region. Subthreshold operation is elegant from the standpoint of the physics and cannot be ignored when subthreshold transistor operation is now 40% of the transistor operating region on recent processes, as shown in Table 1. With the optimal operation point for energy efficiency being near threshold [19], a model that spans operational regions is required. In fact, the work in [19] showed that the supply voltage operating point resulting in the minimum energy per operation is architecture dependent based on the ratio of active devices to leaky devices. This operating point occurs in the perithreshold for modern architectures, thus making a unified and accurate energy model spanning all of these regions critical to energy analysis. Furthermore, the drive for portability has caused a focus on power-constrained computing.

Table 1: Performance summary for nFET devices across successive IBM process nodes.

Node (nm)	V_{DD}	I_{sat} (μA)	V_{thn}	Sub-Vt%	I_{DS0} (pA)	ITRS target $I_{DS0}(pA)$
500	5.0	231	0.64	12.8	<2.5	–
350	3.3	216.3	0.41	12.4	18.2	–
250	2.5	147.5	0.61	24.4	3.7	–
180	1.8	104.1	0.44	24.4	41.9	–
130	1.2	55.5	0.4	33.3	167.1	130
90	1.0	44.4	0.38	38.0	960.3	900
65	1.0	48.5	0.4	40.0	8138.7	6500

Another motivating factor for subthreshold designs is the cost of energy storage. Figure 2 shows energy densities for different chemical energies without considering conversion loss [34]. For common applications, the Lithium-Ion battery has the best energy density for rechargeable electronics, but the energy density of batteries has not scaled with the increases power consumption of the transistor, as seen in Table 1.

In the ideal-case for transistor scaling, the power required to "switch" the transistor will decrease proportionally with the transistor size. The reality of transistor scaling is that the decrease in power is the power of actively changing state, because as the feature sizes have become smaller, the amount of other energy loss has increased. In order to get advantages from smaller transistors, wire distances, contact sizes and oxide thicknesses have decreased as well. These scaling factors have caused an increase in capacitance, which results in more energy being required to change state besides the energy required to charge the gate capacitance of the scaled transistor. Simply put, if the gate capacitance decreases by half, one does not achieve a power savings of half. The result of scaling is indeed more transistors per die area; however, in power-constrained environments, the transistor density has offset the gains of scaling regarding power consumption. This is because the scaling by half does not decrease power consumption by half. Figure 3 illustrates the change in energy storage per kilogram and the number of transistors. The energy requirements have

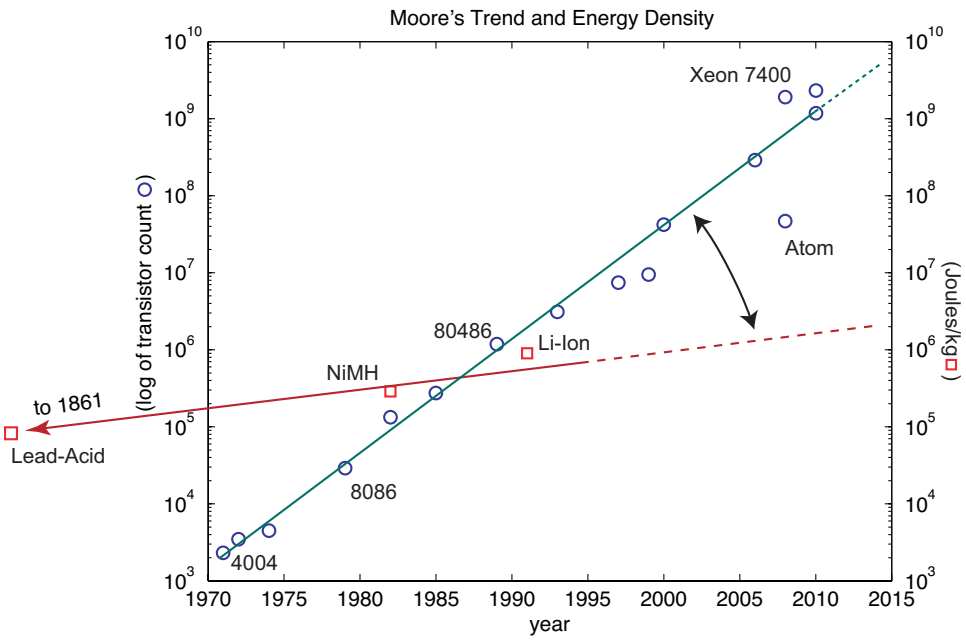


Figure 3: The transistor count of processor products by the Intel Corporation and the energy density of batteries are shown to demonstrate the difference in the rate of total power need verses the realities of energy storage.

increased and the transistor count has increased due the non-ideal electrical scaling behavior, such as the terms in Table 1, and the change in energy storage has been relatively unchanged per weight. If the batteries are not improving at the same rate of processing needs, interesting opportunities exist to fit circuits to power availability, instead of fitting batteries to power needs.

1.2 Squaring up with the Square Law

The real-world is unforgiving, and the physics of transistors are dominated by higher-order effects in modern processes. For this reason, the behavior of a circuit is often simulated as a first step in design. I believe that substantial value still exists in hand-analysis as it gives the designer a solid understanding of the circuit behavior, even if it is not perfectly accurate. Transistors in the digital realm are abstracted to be switches that are "on" or "off", which are the "1"'s and "0"'s of the digital world. As transistor feature sizes have decreased, the supply voltages have decreased, but the

threshold voltage of the devices have not scaled resulting in the subthreshold region of operation to become a more significant percentage of total operation. The "1" and "0" are effectively corner-cases in an operating regime, where in the case of an nFET, a gate voltage of V_{dd} gives the best-case "On" current and a gate voltage of V_{bulk} gives the best-case "Off" current in a CMOS process. The reality in a modern process is that the "Square Law" does not hold due to higher-order effects, and calculating the exact "On" current is a complex process, but the operational corner cases can be used to initially optimize digital designs before simulation [51, 61]. Value exists in enabling the digital designer to find the corner case conditions independent of operating region and V_{dd} through a unified transistor modeling approach.

CHAPTER II

FLOATING-GATE TRANSISTORS

The floating-gate transistor is a device where the gate polysilicon is electrically isolated by oxide allowing the persistent storage of charge. A possible layout approach for this device is illustrated in Figure 4. The "floating-gate" is made more negative through hot-electron injection and more positive through electron tunneling, and its behavior has been well characterized [20]. The floating node allows for persistent charge storage, at the cost of capacitive coupling into the floating node, as seen in

$$\begin{aligned} V_{fg} &= V_Q + \frac{C_{in}}{C_T}V_g + \frac{C_{tun}}{C_T}V_{tun} \\ &+ \frac{C_{gd}}{C_T}V_d + \frac{C_{gs}}{C_T}V_s + \frac{C_{ox}}{C_T}V_b, \end{aligned} \tag{1}$$

where V_{fg} is the floating gate voltage and V_Q represents the voltage offset due to stored charge.

2.1 Electron Tunneling

Electrons are removed from the floating gate via the quantum effect of tunneling through the tunneling capacitor that is explicitly included for this purpose, as illustrated in Figure 4. This process makes the floating node more positive. This tunneling process may be understood as a silicon-dioxide conduction band distortion, as illustrated in Figure 5. Because of the uncertainty of an unobserved electron's position, an electron at one end of the thin silicon-dioxide barrier shown in Figure 5(a) has a finite probability of traversing the capacitor's dielectric even though it cannot do sufficient work to breach the energy barrier [33]. However, with no voltage across the oxide, the oxide thickness must be extremely small for any substantial tunneling through the oxide. However, a voltage drop across the oxide effectively thins the barrier and

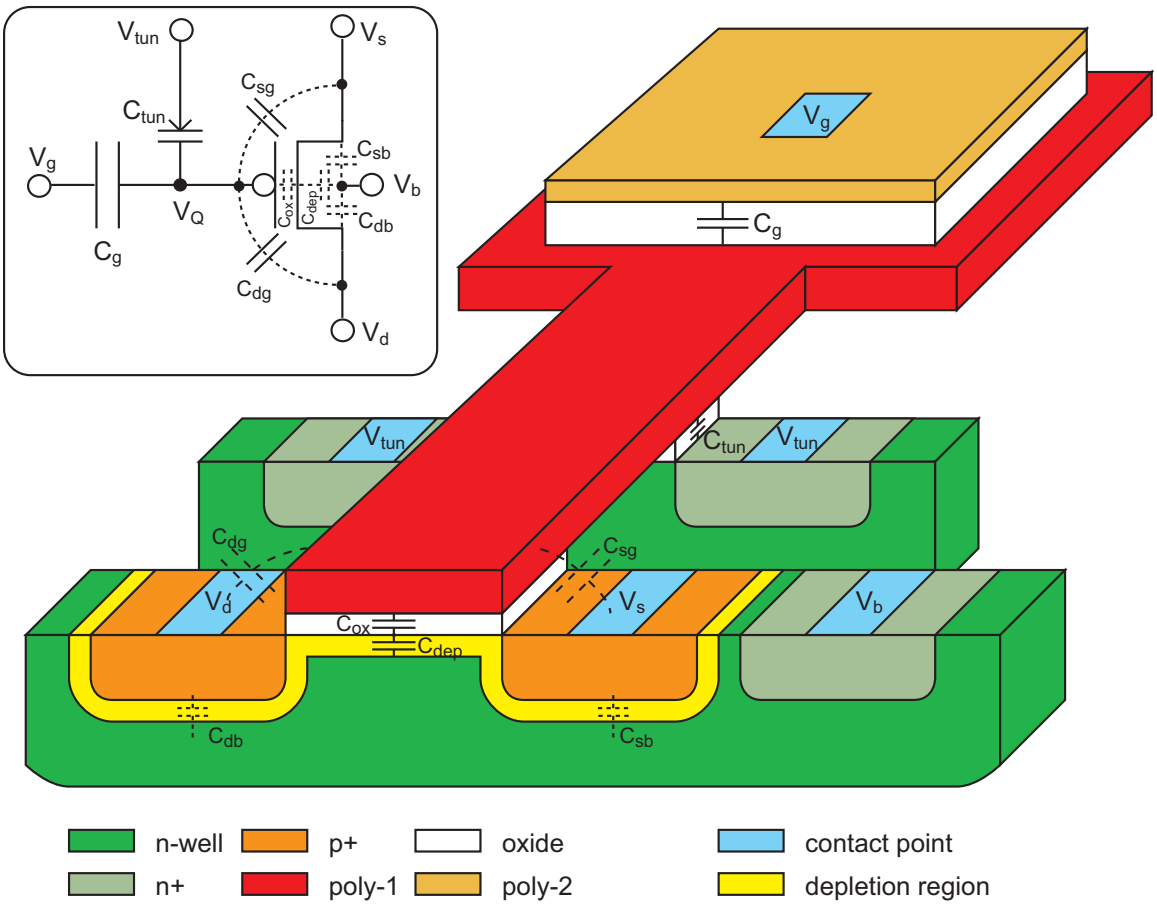


Figure 4: An illustration of a double-polysilicon floating-gate transistor.

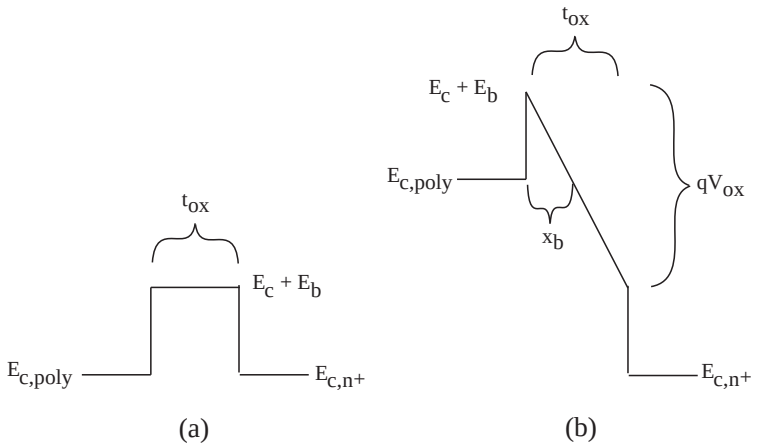


Figure 5: The energy barrier of SiO_2 in a silicon MOSFET (or MOS capacitor) is illustrated. (a) shows the energy barrier with no applied voltage across the oxide, V_{ox}. (b) The effective barrier thickness, x_b, is lowered when a voltage V_{ox} is applied.

exponentially increases the probability that a low-energy electron will traverse it. As shown in Figure 5(b), the energy barrier's new thickness, x_b, is less than the oxide's physical thickness, t_{ox}, because the electron energy has been reduced on the far side of the barrier.

2.2 Subthreshold Injection

Injection in subthreshold is used to make the floating gate node more negative. In the condition of high field at the drain to channel edge, the energy is enough to cause some electrons to leave the conduction band in the nFET, travel through the gate oxide and settle in the floating gate [20]. In the case of the pFET, a similar effect with holes and the valence band will cause an impact ionization that may impart enough energy in the resulting electron to reach the floating gate through the oxide [10].

To add electrons onto the floating gate, as shown in Figure 7, the electrons are "heated" and injected through the oxide. If the pFET is biased with a large voltage drop at the drain (i.e., it is highly saturated) and the channel current is kept below threshold, the high electric field present in the drain-to-channel depletion region imparts high kinetic energy to holes, as shown in Figure 6. The electric field must

surpass a critical value, $\mathcal{E}_{crit}$, in order to drive holes away from the valence band edge in spite of the restoring force of optical phonon emissions and lower energy scattering events. Holes entering into the depletion region have been shown [10] to encounter this critical electric field on average at a critical distance, z_{crit}, after leaving the channel edge.

$$\mathcal{E}_{crit} = \frac{E_R}{q\lambda} \tag{2}$$

$$z_{crit} = \frac{E_R}{\lambda}\frac{\epsilon_{Si}}{q^2 N_d} \tag{3}$$

On average holes become "hot" at $z > z_{crit}$. (z is the distance of carrier flow along the channel and is zero at the channel edge. E_R is the average energy of an optical phonon in silicon, λ is the mean-free length between optical phonon collisions, ϵ_{Si} is the permitivity of silicon, q is an electron charge, and N_d is the substrate doping concentration.) As shown in Figure 6, some of these hot holes collide with lattice sites to ionize the atoms upon impact and generate new electron-hole pairs. Impact-ionization scattering populates an otherwise desolate pFET conduction band with free electrons. Upon return to the channel, some electrons will avoid impact ionization and heat up sufficiently to breach the silicon-dioxide barrier that insulates the gate from the channel.

2.3 Above-threshold Injection

Hot-electron injection thrives in the subthreshold, and exists to a lesser extent in above threshold operation because the drain-to-channel potential barrier is lower per distance, and thereby imparts less energy on the carriers in the channel. However, the energy of the carriers can be increased to promote injection by either shortening the channel or increasing the drain voltage. In Figure 7(a), the drain and source are fixed while the channel inversion is increased from deep subthreshold (depletion) through moderate inversion (around threshold) and strong inversion (above threshold) until

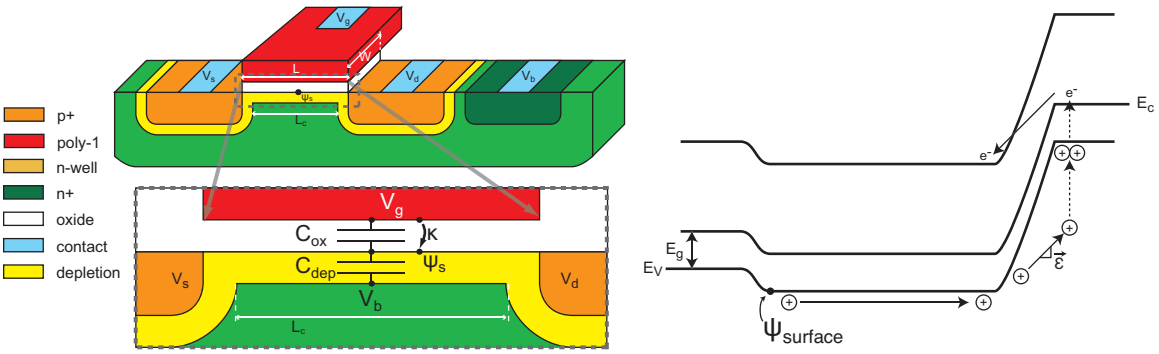


Figure 6: In a pFET below threshold, holes move along an essentially flat valence-band edge through the channel. A large electric field (represented pictorially by the steep slope of the band edges) in the drain-to-channel depletion region empowers the holes to gain sufficient kinetic energy to ionize a lattice site upon impact. The free electron is accelerated sufficiently through the large potential drop that it breaches the oxide and enters the floating gate.

the overdrive voltage is so great the device migrates out of saturation. As the valence band edge rises in the channel near the drain, the electric potential energy gap between the drain node and the source edge of the channel no longer is fully transferred into the drain-to-channel depletion region, and the electric field in the depletion region accelerates carriers at a diminished rate. The electric field is visibly weaker in Fig. 7(a), as seen as a slightly shallower slope of the band edge curves. Furthermore, the built-in voltage of the drain-to-channel junction, Φ_{DC}, is reduced by $\Delta\Phi_{DC}$, which is approximately

$$V_{overdrive} = \kappa(V_g - V_{th}) - V_s, \tag{4}$$

where V_g and V_s are a measure of the gate and source respectively that are *below* the bulk of V_{DD}. The fraction of the gate voltage that is capacitively coupled into the channel is κ. Subsequently, the length of the p-n junction is also reduced:

$$\Delta z_L = \sqrt{\frac{2\epsilon_{Si}\Delta\Phi_{DC}}{qN_d}} \tag{5}$$

Graphically, the overdrive voltage is represented in Figure 7(a) as roughly $q\Phi_3 - q\Phi_1$. As the channel current increases above threshold, the electric field increases in

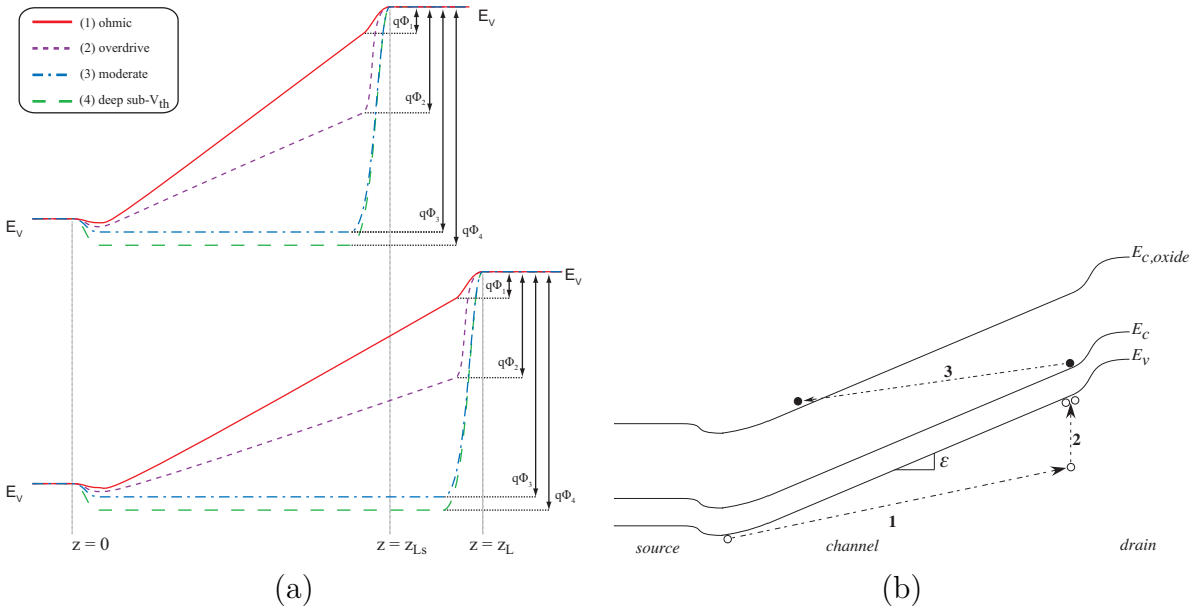


Figure 7: The valence bands are illustrated in (a) and (b). The valence band edges in a short-channel pFET (top set of band diagrams) are compared to those in a longer-channel pFET (bottom set) for fixed drain and source voltages. For each set, the bottom diagram represents a device with little or no channel inversion: the built-in voltage in the p-n junction between the drain and the channel, Φ_4, is greater than the voltage drop between the drain and source. As one progresses up the valence-band edge curves through moderate and strong inversion, the potential drop within the depletion region decreases, implying a subsequent decrease in electric field through the p-n junction. When the device is driven into the ohmic region (the top curve of each set), the potential drop, $q\Phi_1$, is near its minimum possible. Meanwhile, the electric field in the channel (the slope of the band edge) is maximized. In a pFET above threshold, holes move along the rising valence band edge through the channel in step **1**. In the illustration shown in (b), the pFET overdrive accelerates the holes through an electric field greater than the integrated energy dissipation through the channel due to lattice collisions (mostly due to optical phonons). An electric field of greater than 10 V/μm is required to heat the holes in this way, enabling impact ionization in step **2**. [21] Some of the newly generated electrons in the conduction band will become hot enough to inject into the floating gate in step **3**.

the channel but decreases at the channel-drain interface. As a result, hot-electron injection efficiency decreases from the peak values below $I_d = I_{th}$ [21]. However, in favorable device conditions, the critical electric field in (2) may be achieved *in the channel*. It is evident in the top set of bands in Figure 7(a) that the short-channel devices exhibit higher electric fields in the channel during strong inversion.

Consider short-channel pFET biased with a large overdrive voltage and a small drain-to-source difference such that it behaves ohmically as presented in Figure 7(b). In this case, after holes achieve $z > z_{crit}$ in the drain-to-channel depletion region, there is very little distance for them to heat up sufficiently to inject. But consider the channel's electric field, $\mathcal{E}$, which is essentially linear throughout. If the overdrive voltage is great enough that $\mathcal{E} > \mathcal{E}_{crit}$, then the mean holes will heat up through the entire channel's length. Note that in this case, z_{crit}, is the beginning of the channel (on the source side) which I will define as a zero reference point. Similarly, the hole potential at $V(z_{crit})$ can be defined as zero. The average electron's energy at a point in the channel is then

$$E_{h+}(z) = qV(z) - E_R\frac{z}{\lambda}, \tag{6}$$

assuming hole energy high enough to enable maximum optical phonon emission. Consider (6) for the case $z = z_L$ (a hole at the *drain* edge of the channel). The potential rise along the channel length is approximately Φ_{ds}. So, (6) simplifies to

$$E_{h+}(z_L) = q\Phi_{ds} - E_R\frac{z_L}{\lambda}. \tag{7}$$

When the electric field is just *barely* sufficient to induce maximum optical phonon emission along the channel length, it is at $\mathcal{E}_{crit}$. In this case, the average hole energy at any point z along the channel is zero, including at the drain end, and (7) simplifies to

$$\Phi_{ds} = E_R\frac{z_L}{q\lambda}. \tag{8}$$

This is the lower limit of the required source-to-drain voltage drop for a device of

channel length z_L for hot-electron injection to occur in an ohmic device in strong inversion.

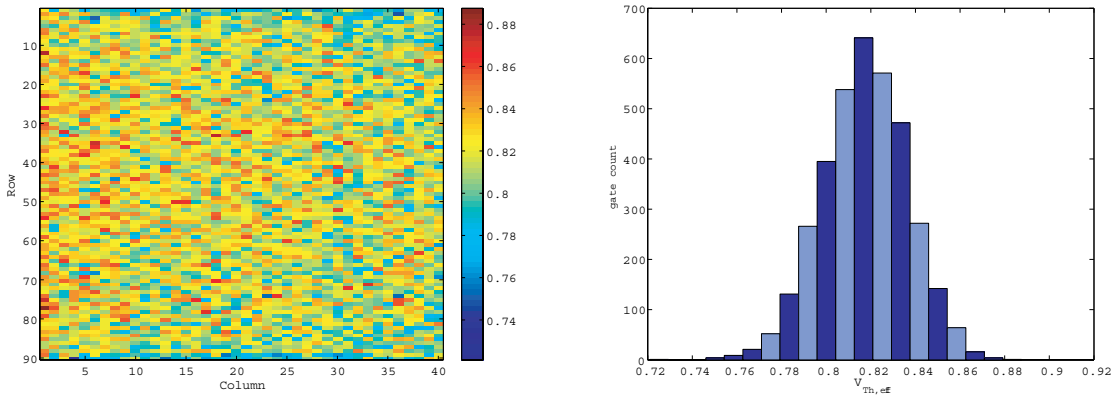
2.4 Non-programmed Floating-Gates

There is a large contrast between digital and analog applications of floating-gates because the charge on the floating node represents more information in analog applications than in digital applications. In the digital world, a large tolerance for the charge on a gate exists because the charge only represents one bit as a "1" or "0", or only a few bits in the case of multi-bit FLASH memory. Therefore, a large voltage tolerance exists for charge representation of the bit state. In the analog domain, the charge on the floating node can represent multiple bits of data, often more than 8-bits. The tendency for floating-gate transistors to accumulate charge in digital systems has been well documented [28]. In the analog domain, the charge accumulation during fabrication on the floating node, V_Q, causes a change in the "back from fab" voltage offset as seen on the floating-gate, V_{fg}. The voltage at the surface of the channel with respect to the voltage at the gate is complex due to the changing capacitance of the channel and depletion region variations. Assuming linear capacitances, the sources of capacitive coupling are

$$\begin{aligned} V_{fg} &= V_Q + \frac{C_{in}}{C_T}V_g + \frac{C_{tun}}{C_T}V_{tun} \\ &+ \frac{C_{gd}}{C_T}V_d + \frac{C_{gs}}{C_T}V_s + \frac{C_{c-ox}}{C_T}V_w, \end{aligned} \tag{9}$$

which are illustrated in Figure 4. The complex oxide divider, C_{c-ox}, is a lumped term that changes with the state of the channel.

The ability to set the "back from fab" state of floating-gate transistors via layout techniques is interesting because it would allow the designer to set the voltage offset on the floating node to a pre-programmed state by either charge removal or changing



(a) (b)

Figure 8: The above are examples of effective threshold voltage (V_{eff}) variance. The offset due of trapped charge on an array of polysilicon floating gates fabricated on a 0.5 µm, SCMOS process available through MOSIS. The effective threshold of the floating-gate varies pre-programming due to trapped charge that remains from fabrication. Plot (a) shows that the effective threshold shift does not have a spacial relationship. Graph (b) shows transistor count in relationship to V_{eff}.

the effective flat-band condition. Rodriguez-Villegas et al. have used stacked metal contacts to the highest metal as a method of normalizing the "back from fab" charge on floating gates with the reasoning that the charge on the polysilicon is normalized due to the flow of metals inhibiting the polysilicon from floating until the final metal etch [47]. Charge leakage due to contacts from metal to polysilicon has been reported [29], and leakage between metal layers has been characterized [50]. However, the impact of this leakage is dependent on usage, and devices such as neuMOS were functional even with a more direct leakage path to the substrate via a switch [52].

To explore trapped charge, two ICs containing floating-gate transistors were fabricated on a 0.5µm, process available through MOSIS. Both ICs used poly1-poly2 capacitors as the gate capacitors to provide linear coupling. A high-density array with 3600 floating-gate pFETs was created for the purpose of exploration of a spacial charge relationship. The other array IC contained 24 floating-gate transistors that

were designed with matching considerations so that the trapped charge could be accessed without the concern of device mismatch. The matched transistors, illustrated in Figure 10, were 18 μm wide and 6 μm long and should have threshold mismatch of less than 2mV of gate voltage offset when referenced to the channel current, I_{ds}, based upon size alone [30]. Additional dummy polysilicon was also included to reduce etch-based mismatch. Given these are floating-gate transistors, matching must also be considered in the capacitive divider that couples into the floating node, shown in Figure 9. Gate poly1-poly2 capacitors for the matched array were approximately 8.6pF to minimize other sources of coupling into the floating node. Furthermore, a n-well below the floating node input gate capacitor followed the gate voltage, so that coupling into the poly1 was from both the poly2 and the n-well. A third source of mismatch is from trapped oxide charge. Since the "back from fab" state was of interest, this charge was not normalized.

Metal contacts were placed on several of the floating gates of the matched array to emulate the methods of Rodriguez et al., but the metal only extended from the polysilicon to the lowest metal. The decision to exclude higher metal layers was made because including contacts to the upper metal layers may not be feasible in layout-dense applications, such as FPAAs. The reasoning behind the metal contacts to polysilicon for charge normalization is the fact that the metal flows will short the gates and equalize the charge [47]; therefore, the lowest metal alone layer may be sufficient to normalize this charge. This is because the metal flows across the whole IC during fabrication and is then etched away. We believed that the same process existed between the top metal layer and the bottom metal layer, which is why the lowest metal layer would be adequate to equalize charge.

The charge on the matched floating-gates was determined by calculating the difference in the gate voltage required for an I_{ds} of 10nA between different devices operating in subthreshold. The subthreshold current equation for a standard pFET

is

$$I_{ds} \;=\; I_0 \mathrm{e}^{\frac{V_{dd}-\kappa V_g+V_s}{U_T}}. \tag{10}$$

One must modify (10) to reflect changes in current due to the capacitive divider caused by the gate capacitor and the tunneling capacitor. Furthermore, it is useful in this case to reference the current down from the threshold current, I_{th}, instead of up from the leakage current, I_0. The capacitances which affect the surface potential are

$$\kappa \;=\; \frac{C_{ox}}{C_{ox}+C_{dep}}, \tag{11}$$

$$V_{fg} \;=\; V_Q + \frac{C_{in}}{C_T}V_g + \frac{C_{tun}}{C_T}V_{tun}, \tag{12}$$

which change Ψ, and thereby affect current matching. (9) gives the gate voltage for the capacitances coupling into the floating node. The pFET device can be made to behave almost identically through layout techniques to match threshold voltages, and by using the same measurement conditions. This allows the terms from (9) to be assumed to be shared between gates, allowing for the only difference to be the stored charge. The threshold prescaler is

$$K \;=\; (\mu_o C_o)\left(\frac{W}{L}\right), \tag{13}$$

which results in

$$I_{th} \;=\; \frac{2K}{\kappa}U_T^2. \tag{14}$$

Starting from the EKV model, I pull out I_{th} into the form seen in (14) [35], the resulting voltage to current relationship is

$$I_{ds} \;=\; I_{th}\mathrm{e}^{\frac{\kappa}{U_T}\left(V_{dd}-V_{fg}+V_{T0}\right)}. \tag{15}$$

Finally, (10) can be rewritten as (15) which references current from I_{th}, assuming that $V_{dd} = V_s$. One can see that for an equal V_g and V_{tun} between devices, V_Q will

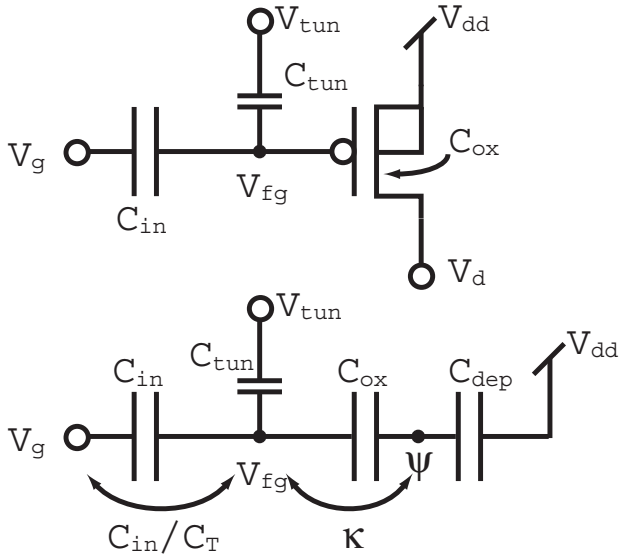


Figure 9: To have matched floating-gate transistors, one must minimize the capacitor mismatch into the floating gate, including that of the capitative divider term, κ. Uniformity of $\frac{C_{in}}{C_T}$ between devices is equally critical as κ matching; otherwise, the surface potential, Ψ, for a given V_g will vary due to process variations.

solely be responsible for any change in I_{ds} for well-matched devices. Comparison is then made by matching I_{ds} for two devices by

$$\Delta V_g \frac{C_{in}}{C_T} \;=\; -\Delta V_Q \;=\; \Delta V_{eff,th}. \tag{16}$$

(16) shows the relationship between the gate voltage and voltage offset due to accumulated charge. Assuming that the capacitive coupling from the gate is well behaved and consistent between devices, as in the case of the matched array, (16) can be further simplified to $\Delta V_g = \Delta V_Q$. Therefore, the difference in the gate voltage required to match currents between the pFET and the floating-gate pFET is the voltage offset due to trapped charge.

A comparison of effective gate voltage due to trapped charge in the high-density, floating-gate pFET (fg-pFET) array is shown in Figure 8(a). The effective gate voltage relationship to gate count can be seen in Figure 8(b). The devices were twice minimum width and length. The fg-pFET devices contained polysilicon gate

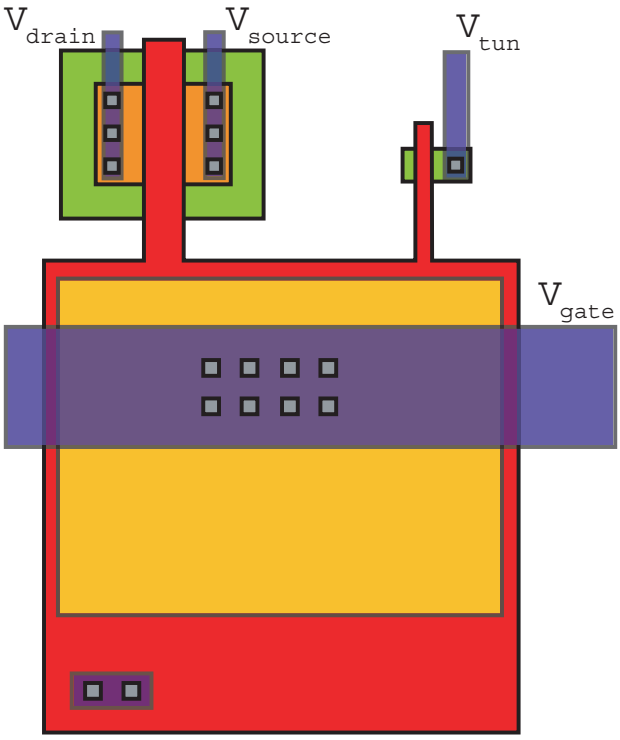


Figure 10: Transistors which were long and wide were designed with matching in mind, including matching orientation and sacrificial poly around the transistors. The channel width is 18 µm and 6µm in length and the draw area of the poly1-poly2 gate cap is approximately 9 picofarads. The tunneling junction is 0.6 µm by 0.6 µm and has negligible coupling. The metal contacts were directly above the polysilicon gate.

capacitors with no metal contacts. The spatial charge distribution was random, and the threshold voltage offset varied by 170mV. The devices were designed for density and not matching; therefore, it is easier to show the a relative offset because uniformity in the flat-band condition cannot be guaranteed between devices.

2.4.1 Matched Transistor Trapped Charge

The array of matched transistors showed excellent matching between transistors on different dies, with a die variance of less than 1mV for the standard pFET transistor. These results are summarized in Table 2. This also shows that mismatch due to trapped charge in the oxide is also negligible. One can then assume that the flat-band condition between devices is matched as well. Therefore it is useful to consider

the stored charge as an offset from standard pFET devices. The range of charge distribution on the polysilicon floating-gate transistors with no contacts was 86mV, and all devices showed negative charge accumulation referenced to the standard pFET device.

Devices with one contact to metal and a drawn metal size of 1.2 μm by 1.2 μm showed charge distribution of 78mV and 76mV, respectively, for two different devices across four ICs. The devices with 10 contacts and a drawn metal size of 29.7 μm by 122.85 μm showed variances of 32mV and 28mV. One can conclude that a contact from polysilicon to the lowest metal does not normalize charge for devices of similar morphologies even if the devices are well-matched between different ICs. More interesting is the fact that charge remained on the floating-nodes even after one year of shelf-time which contrasts previous work [29]. From previous work, one would expect the devices to settle to an offset voltage which satisfies the flat-band condition. Also, the question must be raised about leakage as a function of metal morphology, which possibly why these results differ from the leakage results from St. John et al [50]. It is possible that the flat-band conditions for the 10 contact and single contact

Table 2: The offset due to trapped charge was compared across four ICs as a voltage offset for four devices of identical morphology. Each IC contained two FETs with 1 contact and 10 contacts. The standard pFET devices showed mismatch of less than 1mV between ICs. The contacts on the polysilicon floating-gate to the lowest metal decreased the total variance of charge between devices on the same die and between other dies on the same fabrication run when compared to the polysilicon floating gates with no contacts; however, the data shows clearly that contacts to the lowest metal did not normalize the charge between devices on the same die.

V_{eff}(mV)	IC1	IC2	IC3	IC4	$\lvert\Delta\rvert$
pfet	0	0	0	0	0
fg-pfet	-159	-170	-98	-84	86
10 contacts	-50	-43	-38	-18	32
10 contacts	-79	-77	-62	-51	28
1 contact	-118	-132	-64	-54	78
1 contact	-133	-140	-79	-64	76

devices are different enough for the voltage disparity to exist, and the devices leaked to flat-band before testing; however, one would expect that the flat-band condition for well-matched devices would be different by a few millivolts at most, which does not explain the difference in trapped charge if leakage from the floating gate exists. The charge retention also allows us to dismiss the idea of designing for an effective flat-band by the use of metal-contacts to the floating gates.

CHAPTER III

TRANSISTOR MODELING FOR LOWER-POWER APPLICATIONS

A model that encapsulates the behavior of both above and below threshold is desirable so that a seamless transition is possible between operational regions, even if the interpolation between regions is not representative of the physics.

3.1 Unified Transistor Modeling

This section develops models for transistor behavior independent of transistor operating point. This work is a bulk-referenced model, which accounts well for threshold shifts. Previous work involves modeling the threshold shift as a function of the offset of the source from the threshold, such as the V_{GS} term in the classical digital inverter description. Furthermore, the condition of saturation is defined because one can scale voltage until the saturation assumption no longer holds. Classical texts commonly use the term V_{GS} to describe the gate voltage as it relates to transistor current operation, and most texts neglect that it is a source referenced model where $V_t = V_{t0}(V_{SB})$ [16]. The reality is that the $V_{GS} = V_G - V_S$ assumption does not hold in modern processes. Specifically, V_{GS} does not include the channel divider, and this simplification has not held since at least 350 μm for a source-referenced model, and has never held for a bulk-referenced model. In this section, I will establish an approximation that is more nearly correct for the current relationship for all modes of operation.

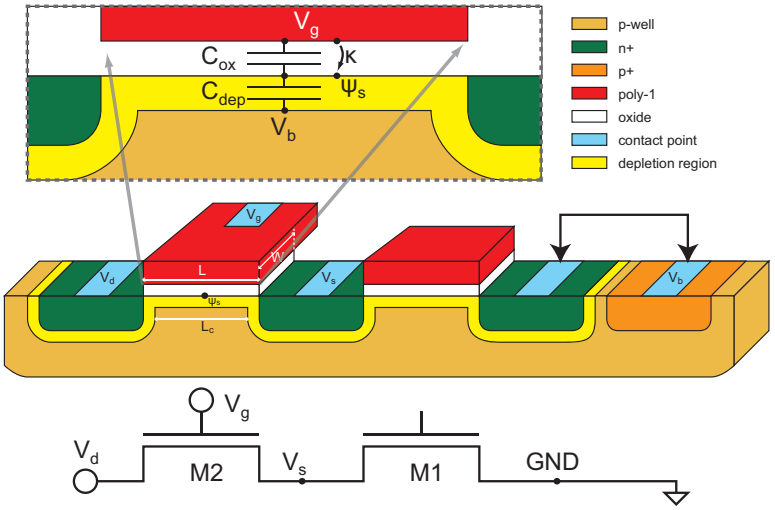


Figure 11: The illustration represents the drawn nFET layout, terminal voltages, and the effect of these voltages on the surface potential, ψ_s. Assuming that both C_{ox} and C_{dep} are fixed, the coupling from the gate voltage to the surface potential can be described as $\psi_s = \kappa V_g$. This also shows the falsity of a V_{GS} term in subthreshold because the gate coupling is independent of the source voltage. For example, in the illustration, the source of M2 is not necessarily at the same potential as the bulk.

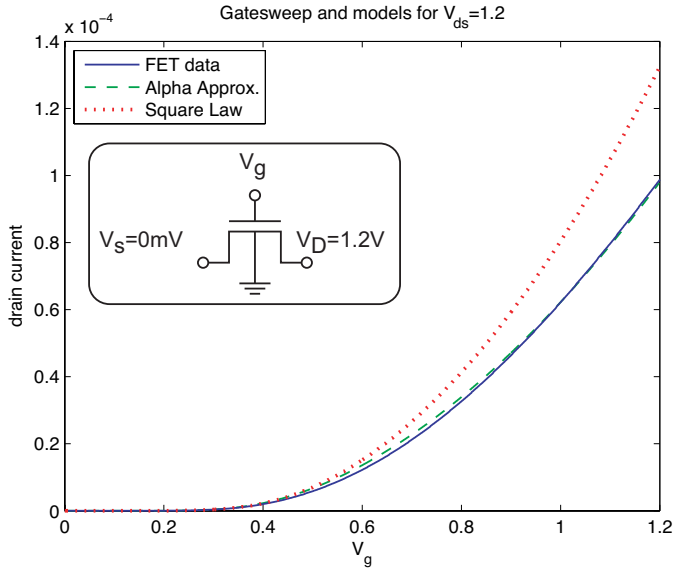


Figure 12: The gatesweep of a 130nm nFET from a commercial process is shown on a linear current scale. The fit of the data compared to the simple model can be much improved by changing only the power term.

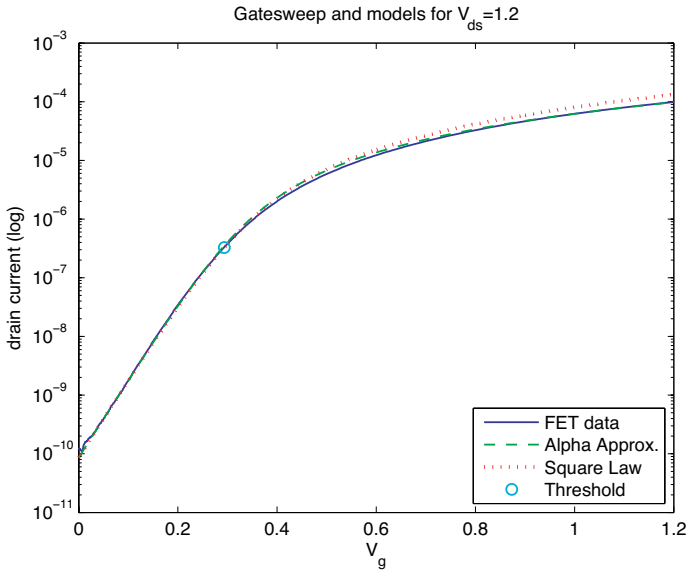


Figure 13: The gatesweep of a 130nm nFET from a commercial process is shown on a logarithmic current scale. Even with the modification of the power term, the simple fit matches device behavior in subthreshold. Furthermore, for a fixed κ from below to above threshold, the model still show a good fit, suggesting that the diffusion encroachment is significant and κ changes only very sightly due to high-order effects in drift movement at this process.

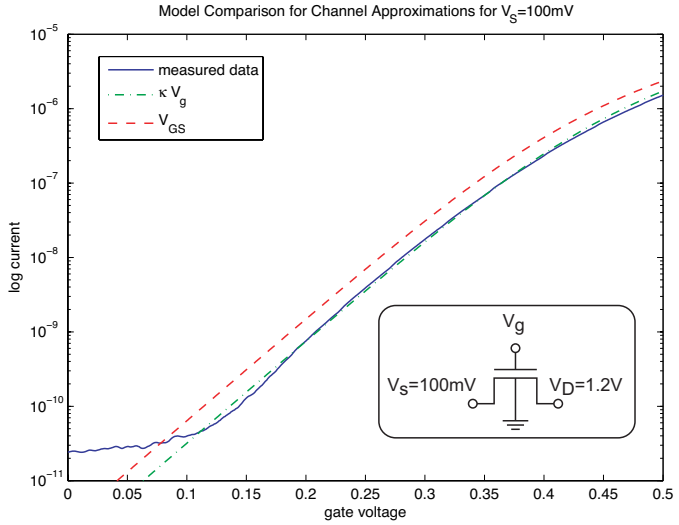


Figure 14: A resistance measurement shows the effect of using the V_{GS} instead of κV_{gb} for the model when $V_S = 100mV$ with a bulk reference of zero volts.

3.1.1 Rethinking V_{GS} for Subthreshold Operation

As processes have become smaller in feature size, the subthreshold regime has moved from a "leakage" condition to a processing condition; however, this operation of devices in the subthreshold regime is well understood, but inaccurately described in reference texts when the source terminal is not tied to the bulk [16]. This inaccuracy exists from the Grey and Meyer text, which is otherwise quite good, to Internet references, such as wikipedia.com[16], and this inaccuracy exists because the model is too simple. The simplifications do not hold in submircon processes. The common approximation for subthreshold operation of the nFET for a device in saturation is

$$I_D \approx I_{Do} e^{\frac{V_{GS}-V_{th}}{\eta U_T}}, \tag{17}$$

where I_{Do} is the current at threshold, U_T is the thermal voltage and η is the "subthreshold slope factor." This slope factor is defined as

$$\eta = 1 + \frac{C_{dep}}{C_{ox}}, \tag{18}$$

where C_{ox} is the oxide capacitance per unit area, and C_{dep} is the depletion capacitance per unit area. The C_{ox} capacitor is a physical capacitor and effectively does not change; whereas, the C_{dep} capacitance changes as the channel inverts. The divider in (18) is correct; however, as it applies to (17) is not true to the physics [17, 42]. Consider the illustration of the nFET channel in Figure 11. The current flow in subthreshold is dominated by diffusion movement, and the barrier is completely dependent on the surface potential, ψ_s. The surface potential is then the related to the gate voltage, V_{gb}, as referenced from the bulk by

$$\psi_s = \kappa V_{gb}, \tag{19}$$

where κ is the capacitive divider to the channel surface through the gate oxide defined as

$$\kappa \equiv \frac{\partial \psi_s}{\partial V_{gb}} = \frac{C_{ox}}{C_{ox} + C_{dep}} = \frac{1}{\eta}. \tag{20}$$

This term is equivalent to the η in (18); however, what one should notice is that the zeroth order analysis of this divider is independent of the source voltage, V_s. Rewriting (17) to reflect this, the resulting current approximation is

$$I_D \approx I_{Do} e^{\frac{\kappa(V_g - V_{th}) - V_s}{U_T}}, \tag{21}$$

which gives a channel divider that is independent of the source voltage. The *slope* in subthreshold becomes $\frac{\kappa}{U_T}$. Of course, the current approximation will change slightly due to DIBL effects, mobility degradation, and the encroachment of the depletion regions [42, 13, 58, 55]; however, (21) is a better approximation than what is given in (17). The C_{dep} capacitance changes as the channel inverts; however, the κ description of the divider to the surface holds well until one approaches threshold [45].

3.1.2 The V_{GS} Condition for Above Threshold Operation

The classical form for above threshold operation is the the square law and for current saturation is

$$I_D \approx \frac{W}{L} \frac{\mu C_{ox}}{2} \left(V_{GS} - V_{th}\right)^2, \tag{22}$$

where the mobility is μ, and μ is constant, disregarding mobility changes with temperature and field [16, 58, 55]. As with the classical subthreshold model, the above threshold model suffers from simplicity. Assuming that the source is fixed, the channel is controlled by the gate voltage, V_g. The depletion capacitor stops changing as the channel enters strong inversion, and one can see in (20) that κ approaches 1 as C_{dep} decreases; however, a capacitance exists due to the depletion regions that exist at the source and drain. For this reason, κ can never actually become completely a function of the oxide capacitor. Modifying (22) to reflect the channel divider results in

$$I_D \approx \frac{W}{L} \frac{\mu C_{ox}}{2\kappa} \left(\kappa \left(V_g - V_{th}\right) - V_s\right)^2, \tag{23}$$

where the C_{ox}/κ is the total capacitance connected to the surface potential along the channel in above threshold.

3.1.3 Unifying Transistor Operation

The Enz-Krummenacher-Vittoz (EKV) model describes the transistor's operation continuously between the subthreshold region of diffusion-based charge movement to the above-threshold region of drift-based charge movement [12]. A variant of the EKV model is the compact EKV model which interpolates around the threshold current [35]. The compact EKV model for a nFET assuming a ground referenced bulk and ignoring the Early voltage, σ, is

$$I_{nFET} = I_f - I_r, \tag{24}$$

$$I_{f,r} = \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{\kappa} \ln^2 \left[1 + e^{\left(\kappa(V_g - V_{T0}) - V_{s,d}\right)/(2U_T)}\right], \tag{25}$$

which combines (21) and (23) into a single equation. The mathematical form of $ln^2\left(1 + e^{\frac{x}{2}}\right)$ between the regions of operation. The compact EKV description for the pFET is

$$I_{pFET} = I_f - I_r, \tag{26}$$

$$I_{thp} = \frac{W}{L} 2U_T^2 \frac{\mu_p C_{ox}}{\kappa}, \tag{27}$$

$$I_{f,r} = I_{thp} \ln^2 \left[1 + e^{\left[\left(\kappa\left(V_b - V_g + V_{thp}\right)\right) - \left(V_b - V_{s,d}\right)\right]/(2U_T)}\right], \tag{28}$$

which is the same as (25) except referenced down from V_{dd}.

3.1.4 Approximating Drift-Current Behavior

The fundamental short-coming of this model is that the "Square Law" from drift movement no longer holds due to higher older effects. Godfrey proposed the mathematical form of

$$f(x) = ln^a(d + e^{bx-c}), \tag{29}$$

in order to encompass these effects[14].

The EKV model is most easily modified for an empirical "α" fit. The form of $ln^{\alpha}\left(1+e^{\frac{x}{\alpha}}\right)$ degrades above threshold region while leaving the subthreshold region untouched. Starting with (25), the resulting modification is

$$I_{f,r} = \frac{W}{L}\alpha U_T^{\alpha}\frac{\mu C_{ox}}{\kappa}\ln^{\alpha}\left[1+e^{\frac{\kappa(V_g-V_{T0})-V_{s,d}}{\alpha U_T}}\right], \tag{30}$$

where α is a fit to the above threshold behavior. Figure 13 illustrates the "Square Law" and α-approximation behavior.

3.1.5 Transistor Saturation

The condition of current saturation is when the drain voltage has little effect on the device current. As one can see in (27), the current through the device is a function of the forward and reverse current, and as the drain voltage increases, the reverse contribution to the current decreases. For the assumption of a saturated drain current, I_D, one may define the condition where the drain current is a percentage of the source current. For a non-inverted channel with V_{ds} greater than $5U_T$, or 125mV, the reverse current is approximately 1% of the channel current and the device can be considered to be in saturation and out of ohmic operation. For a strongly inverted channel, the voltage required for saturation is

$$V_{DSsat} = \kappa\left(V_g - V_{th}\right). \tag{31}$$

3.1.6 Channel Length Modulation

The Early voltage can be modeled by simply adding a σ to the drain voltage in saturation as

$$I_{sat} = I_{thn}\left[\ln^2\left(1+e^{[\kappa(V_g-V_{th})-V_s+\sigma V_d]/(2U_T)}\right)\right], \tag{32}$$

and in ohmic where $V_{ds} < 100mV$ as

$$I_{sat} = I_{thn}\left[\ln^2\left(1+e^{[\kappa(V_g-V_{th})-V_s+\sigma V_d]/(2U_T)}\right)-\ln^2\left(1+e^{[\kappa(V_g-V_{th})-V_d+\sigma V_s]/(2U_T)}\right)\right]. \tag{33}$$

The σ term models the shrinking of the channel due to charge sharing. It has a quantitative measurement that is dependent on the charge sharing at the drain edge. It is worth noting that the value does not have an explicit temperature term. The temperature behavior of σ is explored in Section 5.2.4.

3.2 Energy Delay Model

The energy delay product (EDP) is a metric used to describe the efficiency of digital gates, and any energy efficient technology is effectively seeking to minimize this metric. Much work has been done to model EDP minimization with several limiting assumptions on operating behavior and region [7, 5]. Unifying the operating regions and allowing for scaling of voltages provides a good means for a first-order comparison of EDP across different operating points for the same circuit. The traditional model for delay is based on charging a load through the pFET, as in the case of the inverter shown in Figure 15. This can be modeled as

$$T_d = \frac{KC_{load}V_{dd}}{I_{pFET}}, \tag{34}$$

where C_{load} is the first-order capacitance of the inverter input gate, and the K is a delay parameter made of size and process dependent constants and T_d is the delay [7].

I augment this previous work with the transistor model to be independent of operating point via the methods in the EKV model, as well as including leakage due to the pull-down network. The advantage of this approach is that it includes leakage if it is significant and works for both above threshold and subthreshold. The resulting

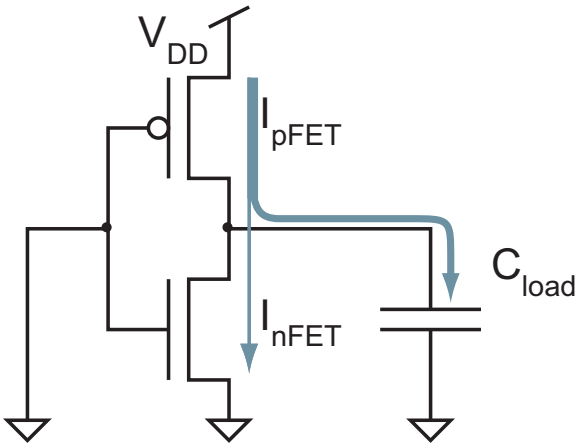


Figure 15: The inverter is generally used as a basic model for current flowing onto a capacitor. I include both pull-down and pull-up networks in the equation model due to the non-negligible subthreshold current that flows directly from V_{dd} to GND.

description for delay is a modified version of (34), resulting in

$$T_d = \frac{KC_{load}V_{dd}}{I_{pFET} - I_{nFET}}, \tag{35}$$

where the current description is independent of the operating point. In (35), the leakage term is subtracted in the denominator, and therefore T_d increases. If V_{dd} is close to the threshold voltage, which is true at the optimal EDP point, then the increase in T_d is significant. T_d ties closely into energy because energy is modeled as the sum of dynamic energy of charging and discharging the capacitive load, C_{load}, and the energy lost to leakage during the transition time. Mathematically, the energy for a gate-level operation is

$$E_{Total} = \alpha_f C_{load} V_{dd}^2 + I_{leak} V_{dd} t_{cycle}, \tag{36}$$

where α_f is the activity factor and I_{leak} is the current of the nFET, I_{nFET}. In (36), I assume that static energy is the leakage power of the digital gate multiplied by the time it takes to complete a cycle, t_{cycle}. Thus, if one wishes to reduce the total static energy used, one needs to minimize I_{leak} or t_{cycle}. The easiest of these to control for the designer is t_{cycle} because it is based upon architecture and, as discussed previously, successive process nodes yield no reduction in I_{leak}.

With synchronous logic, t_{cycle} is tied to a fixed clock period and not the performance of the circuit. Compared to the synchronous case, the asynchronous t_{cycle} is smaller at $\log(t_{cycle})$. This results in the asynchronous energy delay product for a gate in an arithmetic unit being

$$\begin{aligned} EDP &= E_{Total} T_d \\ &= \frac{KC_{load}\left(\alpha C_{load} V_{dd}^3 + I_{leak} V_{dd} \log(t_{cycle})\right)}{I_{pFET} - I_{nFET}}. \end{aligned} \tag{37}$$

CHAPTER IV

ANALOG PROGRAMMABLE CHARACTERISTICS FOR DIGITAL APPLICATIONS

Digital behavior of circuits is generally not adjustable on an individual component granularity. Floating-gate transistors allow for threshold programmability of individual components allowing for improved switches, power consumption modification, and programmable transistor gain.

4.1 Floating-Gate Inverters

Floating-gate inverters with switching threshold and gain programmability have been fabricated in a standard double-poly 0.5 μm process. The inverters had either a single, floating-gate input that shared charge between the gates of the FETs, or the inverters had an isolated input capacitor for both the pFET and nFET. The input capacitors were created from a polysilicon-oxide-polysilicon structure that had linear coupling between the polysilicon layers. The single, shared floating-gate inverter allowed for a programmable switching threshold, which is novel in digital systems without using layout techniques. The inverter that had individual floating-gate inputs for each FET featured programmable gain.

The shared-input capacitor inverter scheme allows for the possibility of variable threshold line buffers to be adjusted for small changes in capacitance due to design, without a change in $\frac{\beta_n}{\beta_p}$, and with relatively little overhead as the indirect injection pFET can be minimum size even when used with larger transistors.

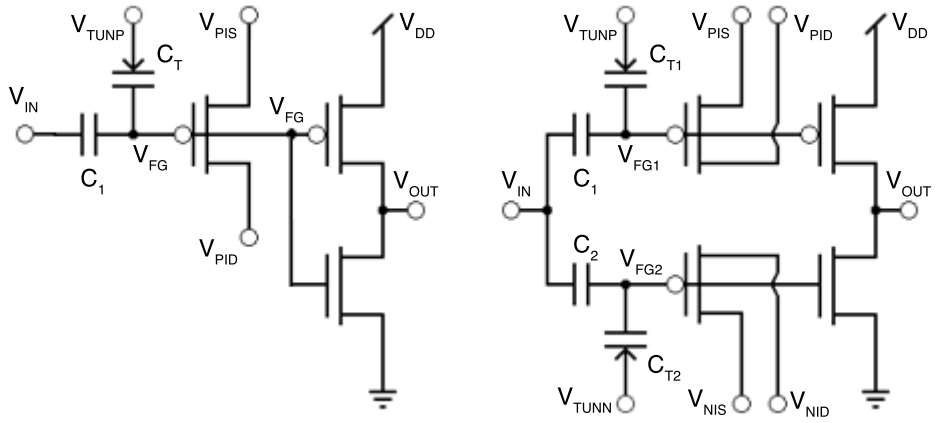
The individually programmable floating-gate inverter also shows promise as a starting point for low-power digital circuits as the programmability allows to the

designer to minimize the current across the device during switching. On the other hand, the power consumption can be increased to overcome capacitance in a circuit in a controlled manner, which is advantageous when trying to achieve a target performance with minimum power. Because switching speed is related to the gain, this type of variable gain circuit could also be used in a precise, programmable manner to compensate for line-skew.

The design presented is an inverter which is divided into two functional blocks: the common CMOS inverter and the analog programming circuitry. Figure 16(a) shows the shared-input floating-gate digital inverter coupled to the analog programming components, which consist of an injection pFET, and floating node with a capacitance (C_{Total}) which consists of a tunneling junction capacitance (C_T) and the input coupling capacitor (C_1) and the capacitance looking into the gate of the transistor . The floating-gates use a dedicated injection pFET in order to create a non-intrusive interface to the "digital" circuit components. Previous methods for injection through a pFET in a complex circuit involved the use of transmission-gates to change current paths for programming. The addition of transmission-gates is intrusive and undesirable for a "digital" circuit as it will affect the overall performance.

4.1.1 Shared-Input Floating-Gate Inverter

The logic threshold of a simple CMOS inverter is set by the ratio of the physical size between the n-type and p-type transistors, ignoring other physical transport effects. The lengths of the transistors are generally the same, and the width of the transistors are generally adjusted by design to determine the logic threshold voltage. In the ideal case where transistors thresholds match and do not change with drain condition, the switching threshold can be determined by the transconductance model where the switching threshold is modeled as a function of electron and hole mobility [54]. For the classic "digital" inverter, this expression can be obtained by equating the currents



(a) (b)

Figure 16: A shared-input floating-gate inverter is illustrated in (a), and a double-input inverter in (b). (a) is the schematic for a single floating-gate inverter with indirect programming circuitry allowing for the non-intrusive addition of a floating node. (b) is the schematic for a dual floating-gate inverter with separate indirect programming circuitry for each transistor.

and solving for the input voltage:

$$\frac{\beta_n}{2}\left(V_{IN} - V_{Tn}\right)^2 \;=\; \frac{\beta_p}{2}\left(V_{DD} - V_{IN} - |V_{Tp}|\right)^2 \tag{38}$$

where V_{IN} is the inverter gate voltage at the midpoint, which is the threshold voltage of the non-floating-gate inverter. This simplifies to

$$V_{IN} \;=\; \frac{V_{DD} - |V_{Tp}| + \sqrt{\frac{\beta_n}{\beta_p}}\left(V_{Tn}\right)}{1 + \sqrt{\frac{\beta_n}{\beta_p}}} \tag{39}$$

where the device ratio, $\frac{\beta_n}{\beta_p}$, determines the value of the midpoint voltage. The floating-gate circuit modifies this value by an offset voltage, V_{charge}, due to the charge on the total capacitance at the floating gate. This modified relationship can be modeled as

$$\Delta V_{fg} \;=\; \frac{C_1}{C_{total}}\Delta V_{IN} + V_{charge}, \tag{40}$$

and then by setting ΔV_{fg} to be equal to ΔV_I, and then solving for ΔV_{IN} yields the modified threshold voltage of the floating-gate inverter as

$$V_{I,modified} \;=\; \frac{C_{total}}{C_1}\left(V_{IN} - V_{charge}\right). \tag{41}$$

Hot electron injection combined with electron tunneling are used to change V_{charge} in order to set the logical threshold. The details of floating-gate programming have been discussed in [22]. In order to inject charge onto the floating node, the voltage potential from V_{PIS} to V_{PID} is raised to promote hot electron injection.

4.1.2 Individual-Input Floating-Gate Inverter

The individual-input floating-gate inverter has a separate floating gate for each transistor, as illustrated in Figure 16(b). Indirect-injection pFETs are present for each floating node for non-intrusive programming of the gates. The individual-input floating-gate inverter adds a degree of complexity over the single floating-gate model with the addition of a second gate capacitor, C_2, as it allows for gain adjustments in above-threshold operation. The inverter can be programmed to operate in either subthreshold or above-threshold modes allowing for power selectivity as well as gain selectivity. It also retains the shared-capacitor inverter's ability for threshold voltage programmability.

The relationship between threshold voltage, V_I, and floating gate charges, $V_{charge,n}$ and $V_{charge,p}$, is derived from the above threshold saturated current equation of

$$I_n = \frac{K_n}{2\kappa_n}\left(\kappa_n(V_{fg,n} - V_T) - V_s\right)^2\left(1 + \frac{V_d}{V_A}\right), \tag{42}$$

where the pFET equation for current is similar but all bias voltages are referenced down from V_{dd}. Setting the two currents equal to each other, I obtain

$$\frac{K_n}{2\kappa_n}\left(\kappa_n(V_{fg,n} - V_{Tn})\right)^2 = \frac{K_p}{2\kappa_p}\left(\kappa_p(V_{dd} - V_{fg,p} - V_{Tp})\right)^2, \tag{43}$$

where $V_{fg,n}$ and $V_{fg,p}$ are the floating gate potentials for the nFET and pFET respectively and are defined as functions of input potential, V_{in}, and programmed floating gate charges, $V_{charge,n}$ and $V_{charge,p}$, as follows:

$$V_{fg,n} = V_{charge,n} + \frac{C_1}{C_{Total}} V_{in},$$
$$V_{fg,p} = V_{charge,p} + \frac{C_2}{C_{Total}} V_{in}.$$

These equations assume that the total capacitance seen at the corresponding floating gates, C_{Total}, are identical. After substituting in the above relationships into (42), assuming that C_2 is equal to C_1, and solving for the value of V_{in} that makes the relationship true, I obtain the value for the threshold voltage as

$$V_{in} = \frac{C_{Total}}{2C_1}(V_{dd} - V_{charge,p} - V_{Tp} - V_{charge,n} - V_{Tn}). \tag{44}$$

The derivation of the gain of the double-capacitor inverter circuit as a function of programmed charges starts with the calculation of g_m from (44). The gain calculation becomes

$$g_m = \frac{\partial I_n}{\partial V_g} = \frac{K_n}{\kappa}\left(\kappa(V_g - V_T) - V_s\right)\left(1 + \frac{V_d}{V_A}\right).$$

For a simple CMOS inverter, the source voltage is at the substrate, and can therefore be set to zero in the above relationship for the nFET and pFET cases. Using the following definitions for the over drive voltages of $V_{on,n}$ and $V_{on,p}$ for the nFET and pFET respectively, $g_{m,n}$ and $g_{m,p}$ can be defined as follows:

$$V_{on,n} = V_{charge,n} + \frac{C_1}{C_{Total}} V_{in} - V_{Tn} \tag{45}$$

$$V_{on,p} = V_{dd} - (V_{charge,p} + \frac{C_1}{C_{Total}} V_{in}) - V_{Tp} \tag{46}$$

$$g_{m,n} = \frac{2I_n}{(V_{on,n})} \tag{47}$$

$$g_{m,p} = \frac{2I_p}{(V_{on,p})} \tag{48}$$

Defining $r_{o,n}$ and $r_{o,p}$ as follows:

$$r_{o,n} = \frac{\partial V_d}{\partial I_n} = \frac{V_{An}}{I_n}, \quad r_{o,p} = \frac{\partial V_d}{\partial I_p} = \frac{V_{Ap}}{I_p} \tag{49}$$

The gain can then be expressed in terms of the $g_{m,n}$, $g_{m,p}$, $r_{o,n}$, and $r_{o,p}$ as,

$$A_v = -(g_{m,n} + g_{m,p})(r_{o,n}//r_{o,p}), \tag{50}$$

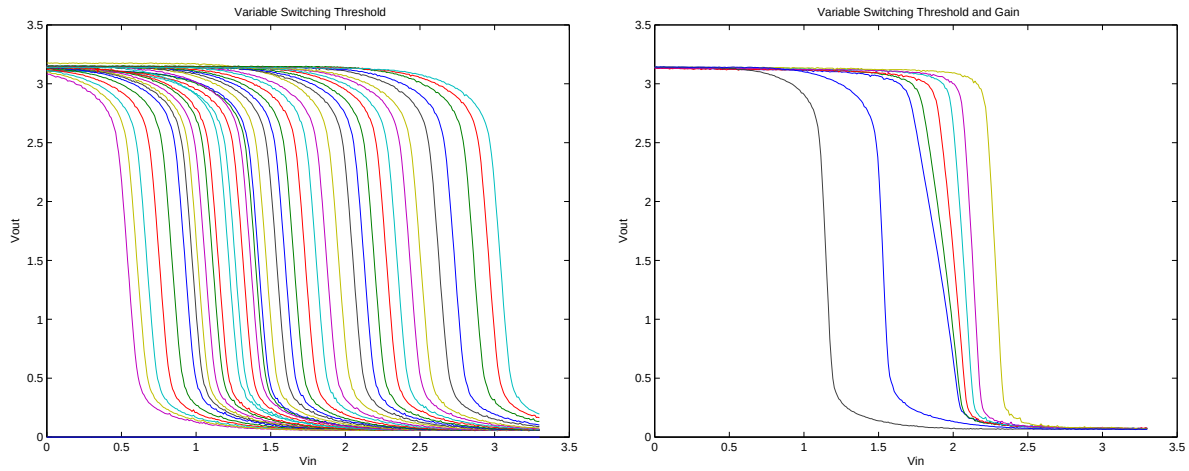
which can be manipulated to the following relationship by substituting (47) and (48) resulting in

$$A_v = -\frac{V_A}{V_{on,n}//V_{on,p}}. \tag{51}$$

Using the above relationship and equations (45) and (46), it can be seen that the gain can be set as a function of the programmed floating gate charges. Increasing $V_{charge,n}$ and decreasing $V_{charge,p}$ would decrease the magnitude of the gain.

4.1.3 Measured Inverter Behavior

The floating-gate inverters were fabricated using a commercially available, double-poly 0.5 μm process. The W/L aspect ratios of the transistors were sized for roughly equal rise and fall times on this process, thus the pFET width is twice that of the nFET. The indirect programming pFET also had a width twice that of minimum size. The indirect programming pFET had the source and drain tied to the positive rail during operation to cut off rogue charge injection, and the tunneling junction was tied to ground. The shared-input floating-gate inverter showed a wide range of threshold programmability, as seen in Figure 17(a). The indirect programming allowed for precise programming as the current through the device was measured from V_{PIS} to V_{PID} without interfering with the design of the inverter. The current



(a) (b)

Figure 17: A shared-input floating-gate inverter is illustrated in (a), and a individually-programmable input inverter in (b). (a) is the schematic for a single floating-gate inverter with indirect programming circuitry allowing for the non-intrusive addition of a floating node. (b) reports data for switching threshold and targeted gain programmability of the dual-capacitor model showing both "digital" inverter and analog inverting amplifier behaviors as the gain is adjusted around an offset threshold.

measured through the indirect programming pFET was used to determine the charge on the gate capacitor, and this current was then correlated to precisely reprogram the threshold to the desired level.

The individually-programmable floating-gate inverter showed a "digital" behavior of wide-range threshold programmability, as well as the analog behavior of a variable gain inverting amplifier. With the threshold-current correlation data from the single capacitor inverter, the dual capacitor inverter was programmed to several chosen threshold offsets, and then the gain was decreased, as seen in Figure 17(b). The gain of the inverting amplifier was then changed in a controlled manner, showing a good degree of precision and range in programmability. The measured gain of the inverter ranged from 100 to under 1.

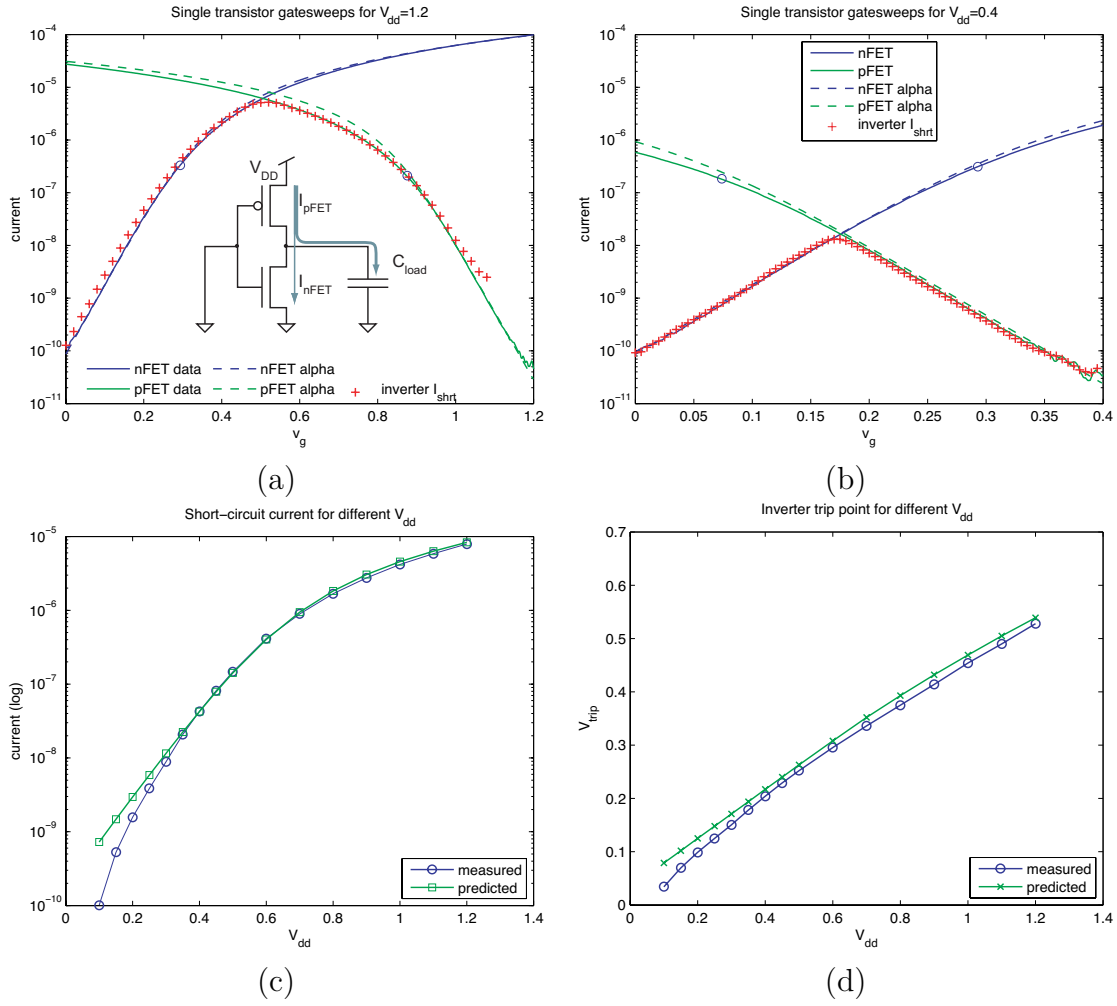


Figure 18: The inverter is generally used as a basic model for digital behavior. Single transistor gatesweeps combined with the measured short-circuit current for an inverter made of similar devices are shown for above and below threshold operation in (a) and (b). The alpha approximation fits are also shown, and the extracted value for alpha was 0.78. The measured inverter short-circuit current compared to the maximum short-circuit current from the unified model are shown in (c) with the measured inverter trip-point compared to the projected trip-point are shown in (d). The model showed good behavior until operation at 20% of the specified process voltage which is most likely due to the model not taking threshold shifts into account.

4.2 Alpha-Model Behavior for Inverters

The fit for the alpha-modified EKV model is shown in Figure 13 with data measured from a 2 μm by 2 μm nFET transistor fabricated on a commercial 130 nm process. The model showed good zeroth order behavior in above-threshold operation and excellent matching to the transistor behavior in subthreshold. The advantages of using (21) with κV_g instead of (17) with a V_{GS} when $V_S \neq 0V$ is obvious in the offset seen in Figure 12(c). For the comparison in Figure 12(c), 100mV was chosen because that is approximately the boundary of ohmic operation [8]. The fact that the measured devices were large reduced the variations in channel divider [49]. The I_{DS0} of the nFET device in Figure 12 is approximately 100pA and the recent process information in Table 1 shows I_{DS0} to be approaching low nano-amps on recent process nodes.

I was able to measure the short-circuit current and trip-point of an inverter structure across different values for V_{dd} for this process, and the measured and predicted values are shown in Figure 18. The maximum short-circuit current and inverter trip-point are well predicted by the model over a range of operating voltages from subthreshold to above-threshold. The short circuit current can be used to estimate the total energy consumption by assuming capacitance and activity factor.

4.3 Capacitively-Biased Floating-Gate CMOS

In the digital system context, floating-gate transistors allow for programming speed of a digital path by allowing for tuning of the current applied to the path. This allows for slack to be removed from synchronous, digital systems. This approach is a modification for the circuits in Figure 16(b) that results in a ring-oscillator layout shown in Figure 19, resulting in a ring oscillator that has programmable cycle times.

This logic approach has direct applications to asynchronous digital systems because the paths may be programmed for exact speed, which is the theoretical ideal

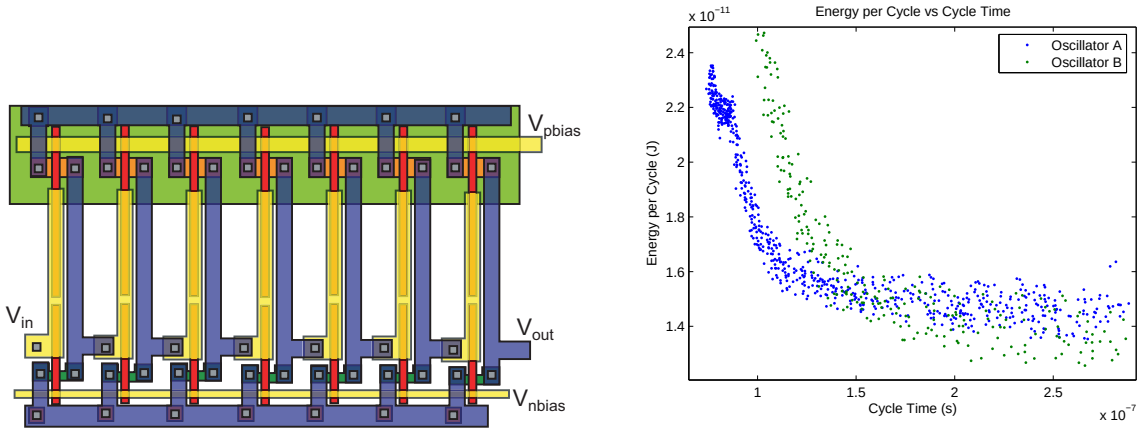


Figure 19: The layout of a capacitively-biased ring oscillator is illustrated in (a), with the change in energy per cycle in (b) for two different oscillator geometries. The ring oscillator was constructed using dual polysilicon capacitors, and changing the bias changed the energy per cycle, as shown in (b).

for an asynchronous system. The voltage seen on the gate is a function of the capacitance values and voltages, as in the following equation, which neglects small offsets due to fringing and overlap capacitance due to gate-source, gate-drain overlaps. These capacitive effects on the gate voltage are described by (52).

$$V_g = V_{in}\frac{C_{in}}{C_{in}+C_{bias}+C_g} + V_{bias}\frac{C_{bias}}{C_{in}+C_{bias}+C_g} + V_{bulk}\frac{C_{bulk}}{C_{in}+C_{bias}+C_g} + V_Q \quad (52)$$

The floating-gate inverter has the benefit that its switching offset can be adjusted by voltage applied to V_{bias}; furthermore, the added capacitance of the floating node in series with the gate capacitance decreases the overall capacitance, which must be charged when switching the gate. However, the gate switches more slowly due to the quadratic relationship of the current to gate voltage, which is why V_{bias} is required if switching speed is increased, at the expense of static power. The actively biased floating-gate inverters individually are difficult to characterize, as the charge moved is small and speed is high; thus, linking these devices into a ring oscillator is appealing, as it allows one to measure the changes to a system. Furthermore, the output is very tangible and thus easy to analyze. The design of the ring oscillators consists of twelve

inverters and one NAND gate that serves as the oscillator-enable pin. By changing to global bias to the nFETs and pFETs of the inverter chain, one can change the oscillation speed of the ring, as well as the power consumed by the devices. The layout and results for the actively biased inverter chain can be seen in Figure 19 (a) and (b).

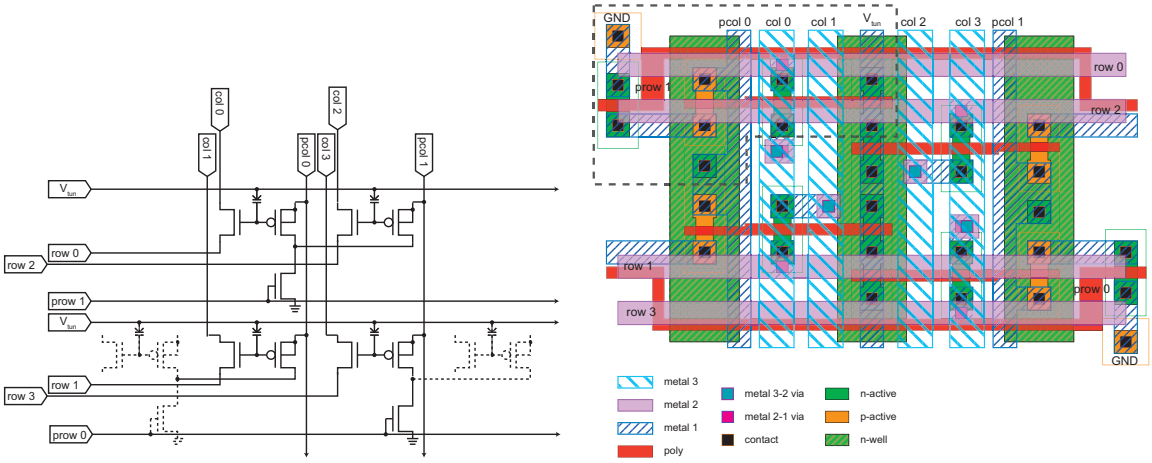
4.4 Single-Poly Crossbar-Matrix for Reprogrammable Architectures

Hot-electron injection is possible in above-threshold operation across a large current range. A switch matrix was fabricated in a 0.5 μm, scalable CMOS process that demonstrated hot-electron injection in short-channel pFET devices. An individual switch, as a resistor, had an "on" resistance between $5k\Omega$ and $9k\Omega$, and an "off" resistance greater $250M\Omega$. Furthermore, the floating-gate device did not have an explicit gate capacitor, allowing for an area savings of over 15% over poly-poly capacitor implementations. The lack of an explicit gate capacitor did not cause adverse behavior of the nFET device as a switch.

4.4.1 Single-Poly Layout Approach

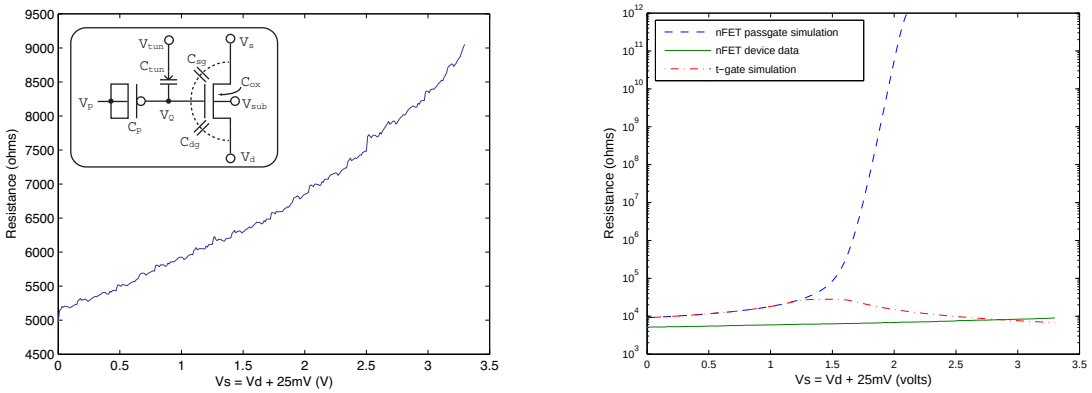
A cross-bar switch matrix was constructed with nFET switches and a floating-gate storage element that is programmed by an injection pFET with no explicit gate capacitor. The injection pFET is fabricated with a channel length of 1.5λ (shorter than the minimum typically permitted) to accommodate hot-electron injection in strong inversion as described in Section 2.3. A nFET is attached as a current cut-off to the drain of the injection pFET to ensure that the undesired injection does not occur because the pFET may act as a short if a large potential exists due its sub-minimum length.

Density of the matrix was further increased by sharing the cut-off transistor between adjacent pFETs. Using a row-column selection scheme based upon the column



(a) (b)

Figure 20: The network schematic is illustrated in (a) with the layout in (b). The crossbar network of fg-nFETs is designed to be tightly tiled to minimize area by sharing the injection enable transistor. The dashed transistors represent transistors in adjacent tiles. Injection of the nFET switch is achieved by setting a large V_{ds} across the pFET via the program column select, *pcol*, and then enabling the row select, *prow*. This method allows for programming in a banked manner. (b) illustrates the layout for the crossbar network. The crossbar network of fg-nFETs is tiled to minimize area by sharing the injection enable transistor. The dashed transistors represent the layout of a single transistor switch. The shared tunneling junction can be seen as a column down the middle of the cell. The lack of an explicit gate cap on the floating-gate allows for area savings as poly-poly capacitors are not required for programming. The total area for 4 switches tiled into a rectangle is $571\mu m^2$ with the savings of $100\mu m^2$ over a poly-poly capacitor implementation. This assumes that the poly1 area required for the poly-poly cap is 5 µm by 5 µm per switch.



(a) (b)

Figure 21: (a) presents measured resistance across a minimum sized nFET with V_{ds} held at 25mV. The increase in resistance is a function of coupling back into the floating node and changes in mobility due to the vertical field. (b) presents the measured 25mV resistance across the fg-nFET passgate plotted with the simulated resistance of non-floating nFET with the gate tied to 3.3 volts, and a transmission gate. The overdriven nFET passgate shows lower resistance than the standard nFET passage, and better resistance than the transmission gate over most of the range for a floating-gate voltage of approximately 7 volts. This shows that designs without an explicit gate capacitor are feasible switch designs.

as injection voltage, and the row as an injection enable, the devices in the matrix can be selected individually. The schematic of the tiled switches and layout are presented in Figure 20(a) and (b).

4.4.2 Switch Performance

Since the digital performance of a device is based upon its impact on the entire system performance, digital system performance is difficult to quantify for a single device; therefore, the characteristics of this device must be compared in an analog sense. The performance measurements were done in the same method as floating-gate pFETs in previous work [15].

The nFET switch has two distinct states as a switch: “on” and “off”. Hot-electron injection was used with 7 volts across the device to force injection over a large current range. The “off” resistance was found experimentally to be greater than $250M\Omega$

without gate control. Tunneling was used to turn the nFET "on". The measured "on" state resistance of the nFET passgate is shown in Figure 21(b), along with other device types. The resistance was measured by fixing the potential across the device, V_{ds}, to be 25mV, and using Ohm's law, resulting in

$$\begin{aligned} V_{ds} = & \left[\tfrac{W\mu C_{cox}}{L2\kappa}\left(2\kappa\left(V_{fg}-V_{T0}\right)\left(V_{ds}\right)+V_s^2-V_d^2\right)\right] \\ 25mV = & \left[\tfrac{W\mu C_{cox}}{L2\kappa}\left(2\kappa\left(V_{fg}-V_{T0}\right)\left(\left(V_s+25mV\right)-V_s\right)+V_s^2-\left(V_s+25mV\right)^2\right)\right]R. \end{aligned} \tag{53}$$

The resistance was calculated with the value of 25mV in order for a comparison with previous work[15, 18]. From (54), one can see that the current will decrease as the source rises, which will result in a change of the effective resistance of the transistor. The "on" resistance of the fg-nFET passgate was found to vary between $5k\Omega$ and $9k\Omega$ for a floating-gate charge, V_Q, of approximately 7 volts. Another change in effective resistance is possible due to the coupling of the source into the floating-node, and the important result of this data is that coupling from the nFET device channel into the floating-gate did not significantly affect the resistance, as seen in Figure 21 (b). This complex capacitance that affects V_{fg} is described as

$$V_{fg} = V_Q + \frac{C_p}{C_T}V_p + \frac{C_{tun}}{C_T}V_{tun} + \frac{C_{gd}}{C_T}V_d + \frac{C_{gs}}{C_T}V_s + \frac{C_{ox}}{C_T}V_{sub} \tag{54}$$

in this architecture, meaning that the condition of the pFET device and tunneling capacitor to ground during runtime dominated coupling into the floating node. The effects of this coupling are clearly seen in Figure 21(a), and further device non-idealities causing the increase in resistance as the channel voltage increases described in other work [8]. Note that no nFET injection was seen on this process.

4.5 Switch Elements

Switch elements are the fundamental building block of reprogrammable architectures, such as FPAAs and FPGAs. Any switch element has two fundamental components,

the signal-path switch and state-storage. Transmission gates are the preferred device for rail-to-rail signal-path operation; however, transmission gates require more physical area and have higher capacitance than single-transistor devices. In the case of a single-transistor switch element in a CMOS process, the designer has a choice of either an nFET or a pFET for the switch element. nFET passgates are used in FPGAs because of their high density, and pFET-based passgates have been used in FPAAs because the gate can be overdriven [15].

Representing the switch matrix of a reprogrammable architecture as a resistor-capacitor network allows for quick analysis of routed paths before spending time on a full-scale SPICE simulation. If assumptions on the bounds passgate operation can be made, a zeroth-order model for resistance through the passgate can be used. The worst-case resistance for an "On" switch is the highest resistance over the operational range. Conversely, the worst-case resistance for an "Off" switch is the lowest resistance. It is desirable to approximate a switch matrix as a RC network for hand calculations; however, the capacitor values for a switch matrix cannot be extracted without knowledge of the matrix architecture, but being able to estimate the switch resistance is a good start. The compact EKV model is a convenient tool for estimating switch resistance because it is a symmetric model that includes both above-threshold and subthreshold operating modes in a single equation.

4.5.1 Theoretical Switch Analysis

The switch model assumes ohmic regime operation, that only V_g, V_s and V_d change and that the degradation of the mobility can be used to approximate higher order terms. It is also assumed that changes in other terms are negligible for above threshold operation. The switch model starts with the compact EKV model, which is a simpler version of the EKV MOSFET model [35, 12]. This model requires extracted values for mobility, the capacitive ratio from the gate to the surface of the channel through

the oxide, κ, and the threshold voltage.

4.5.2 Compact EKV Expansion

Starting from the compact EKV model in (56) with the channel current as the combination of the forward and reverse current in (55) [35], an "On" switch will be in above-threshold operation, ($V_g > V_{T0}$), and the current can be derived as (57). The physical terms that scale the current are drawn width and length of the gate capacitor that are used to calculate C_{ox} to the channel surface. The resulting current equation is

$$I = I_f - I_r \tag{55}$$

$$I_{f,r} = \frac{W}{L} 2U_T^2 \frac{\mu C_{cox}}{\kappa} \ln^2 \left(1 + e^{\left[(V_g - V_{T0}) + (1-\kappa)V_b - V_{s,d}\right]/(2U_T)}\right) \tag{56}$$

$$I_{f,r} = \frac{W}{L} \frac{1}{2} \frac{\mu C_{cox}}{\kappa} \left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_{s,d}\right)^2 . \tag{57}$$

In the context of a nFET switch, one may set the bulk to ground, which simplifies (57) into (58). Furthermore, the ideal switch has a very small voltage drop and therefore drain dependence must be preserved through (55), resulting in (58).

$$I_{f,r} = \frac{W}{L} \frac{1}{2} \frac{\mu C_{cox}}{\kappa} \left(\kappa V_g - \kappa V_{T0} - V_{s,d}\right)^2 \tag{58}$$

$$I_{on} = \frac{W}{L} \frac{\mu C_{cox}}{2\kappa} \left(2\kappa \left(V_g - V_{T0}\right)\left(V_d - V_s\right) + V_s^2 - V_d^2\right) \tag{59}$$

An "Off" switch will be in subthreshold ($V_g < V_{T0}$) and the current can be derived as

$$I_{off} = \frac{W}{L} \frac{2U_T^2 \mu C_{cox} e^{-\kappa V_{T0}/U_T}}{\kappa} \left(e^{(\kappa V_g - V_s)/U_T} - e^{(\kappa V_g - V_d)/U_T}\right), \tag{60}$$

through (56) and (55) by assuming bulk reference.

4.5.3 Mobility Estimation

The channel mobility of electrons will be constant for a particular vertical electric field [58]. As the electric field decreases, the mobility degradation due to interface collisions will also decrease [55]. The assumption is made that the higher- order effects of the MOS capacitor will change the channel current less than fixing the effective mobility as the measured the mobility found at a bulk-referenced source. A test called a "gatesweep" where the gate of a device is swept across the operating range to measure the change in current for change in gate voltage is valuable for determining device characteristics. This gatesweep can be used to approximate the mobility at a V_{ds} resulting in an effective mobility, μ_g, for a V_g. Starting from (59), one can assume that squared terms have a negligible contribution to the channel current, I_{on}, for small V_{ds} where $V_s = 0V$ and therefore can be neglected. This assumption also allows one to cancel the κ terms, which results in an equation independent of the channel coupling term in (61) resulting in

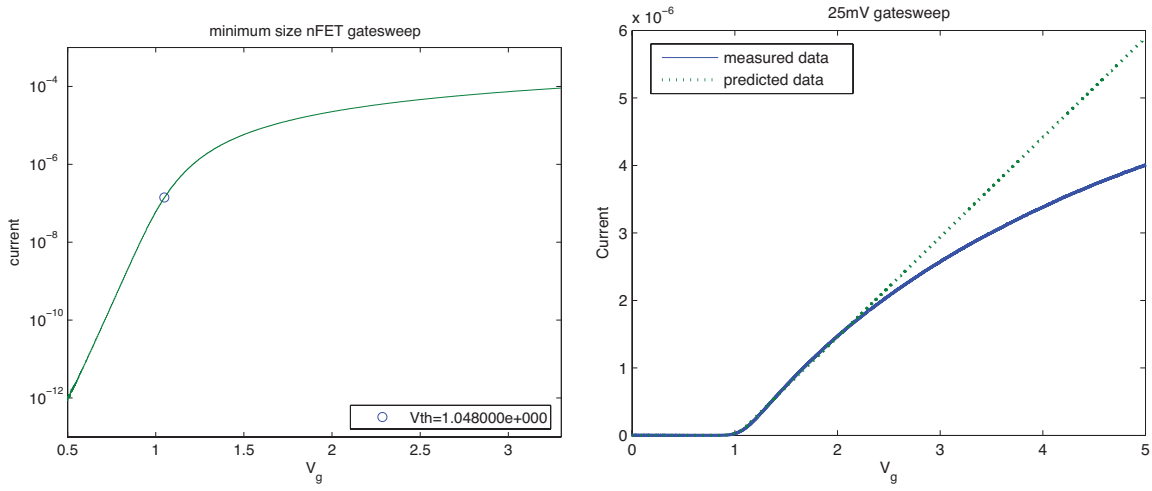
$$I = \frac{W}{L}\mu C_{cox}\left(\left(V_g - V_{T0}\right)\left(V_d - V_s\right)\right). \tag{61}$$

Mobility, μ_g, at specific V_g and V_{ds} can now be solved by rearranging (61), resulting in

$$\mu_g = \frac{L}{W}\frac{I}{C_{ox}\left(\left(V_g - V_{T0}\right)\left(V_d - V_s\right)\right)}. \tag{62}$$

4.5.4 Other parameters

The threshold voltage is defined as the point where the current in the channel is half drift and half diffusion [17]. The nFET threshold is extracted from a gatesweep with the source and drain at GND and V_{dd} where the measured current is half that predicted by the subthreshold current equation. The value of κ above and below threshold is extracted from the same gatesweep [42].



(a) (b)

Figure 22: (a) and (b) report measured data for two different drain conditions. (a) presents a gatesweep for $V_{ds} = 3.3V$ of a minimum-sized, nFET transistor used to extract the above and below threshold κ, and the threshold voltage V_{T0}. The threshold voltage is the gate voltage, V_g, where the measured current is half that predicted by the subthreshold current equation. The body coefficient, κ, was extracted by fitting the compact EKV model to the curve for V_g from 100mV to 1V above V_{T0} so that velocity saturation was avoided. For a V_{ds} for 3.3V, mobility degradation due to velocity saturation was noticed at approximately a V_g of 2.8V. (b) presents a gatesweep of $V_{ds} = 25mV$ where $V_s = 0V$ was used to approximate the low field mobility for $V_g = 3.3V$. The small V_{ds} forces the transistor into the ohmic regime of operation; however, mobility degradation due to high field results in non-linear operation. The mobility of electrons for ohmic regime operation will be limited by mean time between collisions at the oxide interface as they are attracted to the high field from the gate. The mobility is highest for $V_g = V_{T0}$ and degrades steadily as the gate voltage increases.

4.5.5 Switch Resistance

The resistance of a passgate is ideally linear, and can be found by the equation $V = IR$. To calculate the "On" resistance, a substitution of $V_{ds} = V_d - V_s$ is made into (59) to simplify the equation of

$$I_{on} = \frac{W}{L}\frac{\mu C_{cox}}{2\kappa}\left(2\kappa\left(V_g - V_{T0}\right)\left(V_{ds}\right) + \left(V_d - V_{ds}\right)^2 - V_d^2\right). \tag{63}$$

Expanding and simplifying (63) results in

$$I_{on} \;=\; \frac{W}{L}\frac{\mu C_{cox}}{2\kappa}\left(2\kappa\left(V_g - V_{T0}\right)\left(V_{ds}\right) - 2V_{ds}V_d + V_{ds}^2\right). \tag{64}$$

Substituting (64) into $V = IR$ simplifies to (65) yields

$$R_{on} = \frac{V_{ds}}{I_{on}} = \frac{\frac{L}{W}\frac{2\kappa}{\mu_g C_{cox}}}{\left(2\kappa\left(V_g - V_{T0}\right) - V_d - V_s.\right)}. \tag{65}$$

Equation (65) calculates the worst-case "On" resistance for an operating point. Equation (65) is valid as long as the device remains above threshold. At the point where the solution to (59) becomes less than zero, the device enters subthreshold operation and the above-threshold assumption fails. Starting with (59), the minimum gate-voltage, V_g, required for the passgate to remain "On" for a given V_{ds} is defined by (68), which is obtained through

$$0 = \frac{W}{L}\frac{\mu C_{cox}}{2\kappa}\left(2\kappa\left(V_g - V_{T0}\right)\left(V_d - V_s\right) + V_s^2 - V_d^2\right), \tag{66}$$

and then substitution for

$$0 = 2\kappa\left(V_g - V_{T0}\right) - V_d - V_s, \tag{67}$$

finally resulting in

$$V_g = \frac{V_d + V_s}{2\kappa} + V_{T0}. \tag{68}$$

In the case of a passgate, the minimum V_g for rail-to-rail operation would be found by setting $V_d = V_{dd}$. The "Off" resistance of a switch is found through the subthreshold current equation. Starting from (60), the maximum current is the threshold current then becomes

$$I_{off} = \frac{W}{L}\frac{2U_T^2\mu C_{cox}e^{-(\kappa V_{T0})/U_T}}{\kappa}e^{\kappa V_g/U_T}\left(e^{-V_s/U_T} - e^{-V_d/U_T}\right). \tag{69}$$

Assuming the worst-case current when conduction is not desired, the resulting in a relationship is independent of the source and drain voltages because the dominating term are V_g and V_{T0}. Applying $V = IR$ to (69) and solving for resistance results in

$$R_{off} = \frac{L\kappa e^{(\kappa V_{T0}-V_g)/U_T}\left(V_d - V_s\right)}{WU_T^2\mu C_{cox}\left(e^{-V_s/U_T} - e^{-V_d/U_T}\right)}, \tag{70}$$

when $V_{DS} < 4U_T$. (70) simplifies in saturation to be

$$R_{off} = \frac{L\kappa e^{(\kappa V_{T0}-V_g)/U_T} - V_s}{WU_T^2\mu C_{cox}\left(e^{-V_s/U_T}\right)}, \tag{71}$$

The value of R_{off} is useful because a switch matrix has many parallel switches and therefore parallel resistance. This value can be used to estimate the worst-case leakage through the "off" switches in parallel by dividing R_{off} by the number of "off" switches on a wire.

4.5.6 nFET Passgate Analysis

The theoretical analysis presented in Section 4.5.1 was verified through an IC fabricated in a 0.5 µm, scalable CMOS process available through MOSIS. 25mV was fixed across a minimum-sized nFET passgate and swept for different gate voltages. The BSIM 3.5f device information provided by MOSIS was used as a comparison.

4.5.7 Parameter Extraction

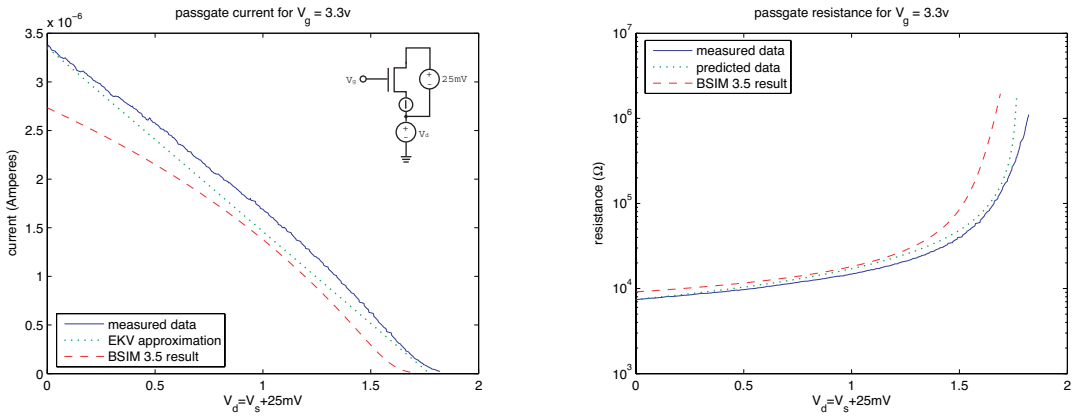
The above threshold current equation (54) requires the above-threshold κ and effective mobility at $V_{ds} = 25mV$. A gatesweep with $V_{ds} = 3.3V$ was used to extract κ and V_{T0}.

A log-fit of the subthreshold region, derived from (56), was used to identify threshold, Figure 22(a). κ was extracted by fitting (54) to the curve for V_g from 100mV to 1V above V_{T0} so that velocity saturation was avoided. For a V_{ds} of 3.3V, mobility degradation due to velocity saturation was noticed at approximately $V_g = 2.8V$ for this particular device. A gatesweep with $V_{ds} = 25mV$, shown in Figure 22(b), was used to extract the low-field mobility using (62) for a gate voltage of 3.3V. The small V_{ds} forces the transistor into the ohmic regime of operation; however, mobility degradation due to high field results in nonlinear operation. The mobility of electrons for ohmic regime operation will be limited by mean time between collisions at the oxide interface as they are attracted to the high field from the gate [58]. The mobility is highest at $V_g = V_{T0}$ and degrades steadily as the gate voltage increases. This mobility degradation effect can clearly be seen in Figure 22(b). The extracted mobility at V_{T0} nearly matched the $U0$ term given as a BSIM parameter by MOSIS.

4.5.8 Applied Model

Using the compact EKV model from Section 4.5.2, and the parameters extracted in Section 4.5.7, the model can now be compared to actual data. The source and drain terminal of a minimum-sized, nFET passgate with a gate voltage of 3.3 volts was swept from GND to V_{dd} with a $25mV$ potential forced between the drain and source. The measured data, the BSIM simulation from the parameters provided by MOSIS and the compact EKV model from Section 4.5.2 are shown in Figure 23(a). Even though higher-order effects are ignored, the model represents a worst-case current value for the device until the device enters subthreshold and then becomes invalid at the point calculated by (68).

The passgate resistance is reported in Figure 23(b). The passgate resistance increases as V_d increases until a worst-case R_{on} resistance of approximately $10k\Omega$ at



(a) (b)

Figure 23: (a) and (b) report the measured passgate behavior as current and resistance respectively. In (a), the source and drain terminal of a minimum-sized nFET passgate with a applied gate voltage of 3.3 volts was swept from GND to V_{dd} with a $25mV$ potential forced between the drain and source. The figures include the results from the measured data, the BSIM simulation from the parameters provided by MOSIS and the compact EKV model. The model describes the worst-case operation of the nFET while it remains in above threshold operation even though it neglects higher-order effects. In (b), the resistance of a minimum-sized nFET passgate with a gate voltage of 3.3 volts. The worst-case resistance is $10k\Omega$ at $V_d = 1.8V$ with a fixed V_{ds} of 25mV. The resistor approximation is valid as long as $V_d < 1.8V$.

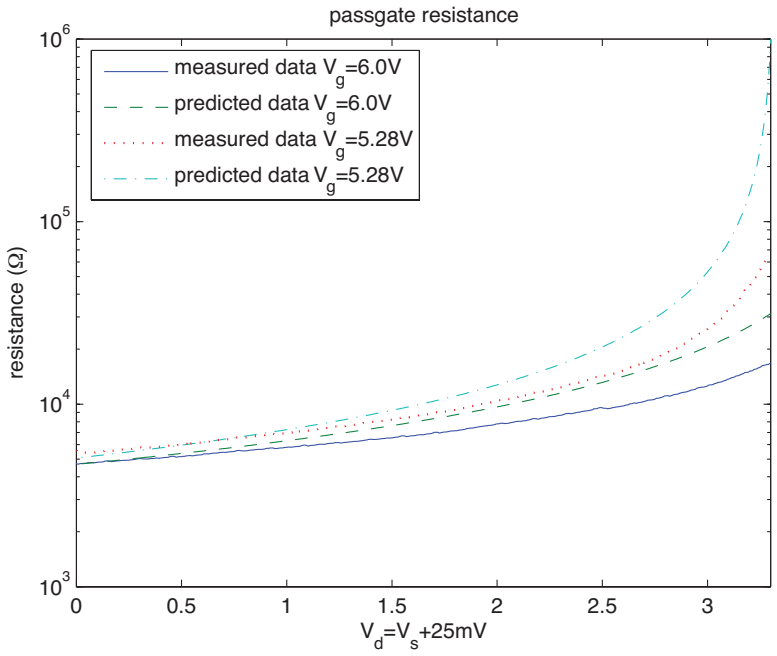


Figure 24: The figure reports the measured resistance of a minimum-sized nFET passgate with the gate voltage, V_g, fixed at 6.0 volts so that the device operates as a "switch" between input voltages from $0V$ and $3.3V$. The worst-case resistance is $3k\Omega$ at $V_d = 3.3V$ with a fixed V_{ds} of 25mV. The minimum V_g required for 3.3V operation was predicted to be 5.28 volts and one can see the resistance increase as V_d approaches $3.3V$.

$V_d = 1.8V$, after which the switch resistance moves to the R_{off} model. This demonstrates that for the device to be described using the resistor model in (65), V_{dd} cannot exceed $1.8V$ if the gate voltage, V_g, is only 3.3 volts, at which point the resistance is approximated by (70). In the case of a passgate-based FPAA switch matrix, this method could be used to determine the maximum signal amplitude about a common-mode voltage.

It is obvious that for the resistor model to hold, V_g must be higher than the maximum V_d seen by the passgate. As an example of the ease of using this method for resistor representation, starting the boundary calculation from (68), I can calculate the minimum V_g required for maximum $V_d = 3.3V$ as 5.28 volts. This boundary condition is the absolute minimum gate voltage required for the above threshold approximation to remain valid. In practice, one would want to exceed this value. Figure 24 shows the measured resistance through the nFET passgate for a V_g of $5.28V$ and $6.0V$. The value of $V_g = 6.0V$ was picked as it was the next whole number above the calculated minimum voltage. The worst-case resistance for $V_g = 6.0V$ predicted by the model is $3k\Omega$ at $V_d = 3.3V$, whereas the the actual resistance value is $1.6k\Omega$. The difference between the predicted and measured values is due to the fact that the second-order MOSFET effects have been completely ignored in the compact EKV model from which this work is based.

4.6 Predicting The Velocity Saturation Boundary of Above Threshold Operation

The saturation current is limited by velocity saturation at submicron feature sizes and, as feature sizes decrease, the velocity saturation effects are observable closer and closer to threshold. I believe that velocity saturation is reached as soon as the device enters above-threshold operation for effective lengths less than 25 nm. I take a qualitative approach to the examination of velocity saturation at subthreshold, above

threshold, and at threshold. In the case of subthreshold operation, the diffusion-based movement of charge results in some number of electrons moving a distance. The velocity in subthreshold is a function of the number of carriers until the channel inverts, resulting in

$$v = \frac{1}{n} D_n \frac{\partial n}{\partial x}, \tag{72}$$

where n is the number of electrons, D_n is the diffusion constant, and x is the effective channel length. This velocity is illustrated in Figure 25(a). If velocity saturation is known, one can solve for what conditions cause saturation. To estimate velocity saturation for carriers in diffusion, I start with (72) and assume that motion across the channel is linear, resulting in

$$v = D_n \frac{n_s}{l} \left(\frac{1}{n_s \left(1 - x/l\right)} \right) = \frac{D_n}{l - x}, \tag{73}$$

where n_s is the carrier density at the source side and l is the channel length. One will quickly notice that as $x \rightarrow l$, the velocity approaches infinity, showing that qualitatively, it is possible to reach $v = v_{sat}$ in subthreshold.

For the above threshold case, I first assume that velocity is a function of field and mobility, and therefore, the current is simply charge moving through the channel at some velocity, as defined by $I = vQl$, where I is current, v is velocity, l is length, and Q is charge that is proportional to the gate voltage as

$$Q \propto \kappa \left(V_g - V_{th}\right) - V_s. \tag{74}$$

The band structure for the resulting case is illustrated in Figure 25(b). For a low source-drain field, a "saturated" state exists that is independent of horizontal field due to increased vertical field that causes increased collisions as the gate voltage increases [8]. The more commonly considered case is for large source-drain potentials and saturation is due to both horizontal and vertical field.

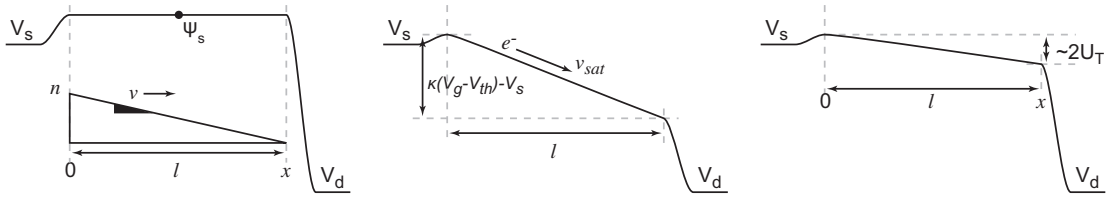


Figure 25: The conduction band is drawn for a nFET in diffusion transport, drift transport and the threshold boundary where charge transport is half drift and half diffusion in (a), (b) and (c) respectively. Velocity saturation is simply a limit of the movement of carriers across the effective channel length. In both diffusion and drift transport, a velocity saturation condition exists. In diffusion transport (subthreshold), the velocity is a function of the number of carriers at the source and the effective channel length, as illustrated in (a). The drift transport (above threshold) case for velocity is simply the channel length and applied voltage, as illustrated in (b). The most interesting case is where channel current is half drift and half diffusion as illustrated in (c). In this case, the voltage is approximately $2U_T$, and the velocity of the carriers may be high enough that velocity saturation is reached just at threshold for short channel lengths.

The most interesting case for velocity saturation is at threshold when drift current is equal to diffusion current, as illustrated in Figure 25(c). The simplest way think about this condition is to consider a length of n-doped silicon. Making the assumption that the charge distribution is linear, I can make a series of simplification for combined drift and diffusion movement and lump terms, resulting in

$$\begin{aligned} J &= q\mu_n n\xi + qD_n\frac{\partial n}{\partial x} \\ J &= q\mu_n n\left[\frac{V}{l}\right] + qD_n\left[\frac{n}{l}\right] \\ J &= \frac{q\mu_n n}{l}\left(V + U_t\right). \end{aligned} \tag{75}$$

In (75), the equation shows that when drift is equal to diffusion, $V = U_t$, resulting in a potential drop across the channel of approximately $2U_t$. What is interesting is that velocity saturation occurs for fields of about 2 V/μm for sub-micron feature size processes. With the assumption that $v_{sat} = 2V/\mu m$, consider slope in the channel

illustrated in Figure 25 (c), the field in this case is

$$\xi = \frac{2U_T}{l}. \tag{76}$$

If (76) is rearranged and the field is replaced with the saturation assumption, the resulting equation for length is

$$l = \frac{2U_T}{v_{sat}} \approx 25nm, \tag{77}$$

showing qualitatively that a device of 25 nm in length is in velocity saturation as soon as one enters above threshold operation. This fact demonstrates the importance of a model that interpolates well through transitional regions and has significant ramifications for digital design because the saturation current will be just at threshold.

CHAPTER V

TEMPERATURE ROBUST SUBTHRESHOLD CIRCUITS THROUGH A BALANCED FORCE APPROACH

The subthreshold region of transistor operation has a strong dependence on the thermal voltage and channel divider; however, implementations of these circuits are surprisingly robust. I present a design method for subthreshold circuits that is intuitive with applications in both teaching and design. The diffusion-based current has simple physics that allows one to make the assumption of a fixed source-channel potential that will then govern the voltages seen across the remaining devices in the circuit.

5.1 Diffusion Movement: A Capacitive Approach

Electrical devices are analyzed by balancing the forces that enter and exit the system. The flux that enters the system must equal the flux that exits. This is analogous to the analysis of a truss through static forces in mechanical engineering. In an electrical system, the balance of force can also be used to analyze a system. MOSFET devices in subthreshold have very simple physics due to diffusion transport, and this device behavior is dominated by the surface potential at the channel. The capacitive dividers that couple from the gate of the device to the channel surface can be used in a "balance of force" approach to model device behavior. In a circuit where symmetry exists, the balance of electrical force is analogous to the balance of mechanical force that one would see in a beam, as illustrated in Figure 26. This approach allows the modeling of device behavior even if it is not completely true to the physics. This balance of force approach can be taken because the current is completely dependent on the source-channel barrier. In the simplest sense, the source-channel barrier can be represented

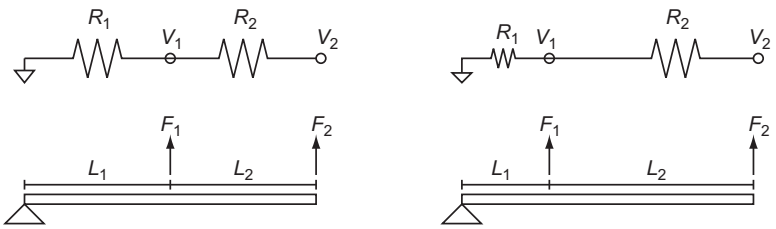


Figure 26: An example of electrical and mechanical systems that can be solved based upon a balanced force approach. The length of a beam is synonymous to resistance, and voltage is synonymous to force.

in the form of

$$I = I_x f\left(\frac{\Phi_{SC}}{U_T}\right), \tag{78}$$

where f is some function behavior, U_T is the thermal voltage, Φ_{SC} is the source to channel potential, and I_x is some prescalar. Behavior of the current is primarily based then the capitative coupling that changes around Φ_{SC}, because it is effectively fixed.

5.1.1 Diffusion Transport

The current though a subthreshold MOSFET is set by the source-channel potential, Φ_{SC}; however, the standard form MOSFET equations are voltages, so I will start from a voltage driven current description. The Enz-Krummenacher-Vittoz (EKV) model describes the transistor's operation continuously between the subthreshold region of diffusion-based charge movement to the above-threshold region of drift-based charge movement [12]. A variant of the EKV model is the compact EKV model, which interpolates around the threshold current and does not include drain effects, such channel modulation or DIBL[35] . The compact EKV model for a nFET is

$$I = I_f - I_r, \tag{79}$$

$$I_{f,r} = \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{\kappa} \ln^2\left(1 + e^{\left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_{s,d}\right)/2U_T}\right), \tag{80}$$

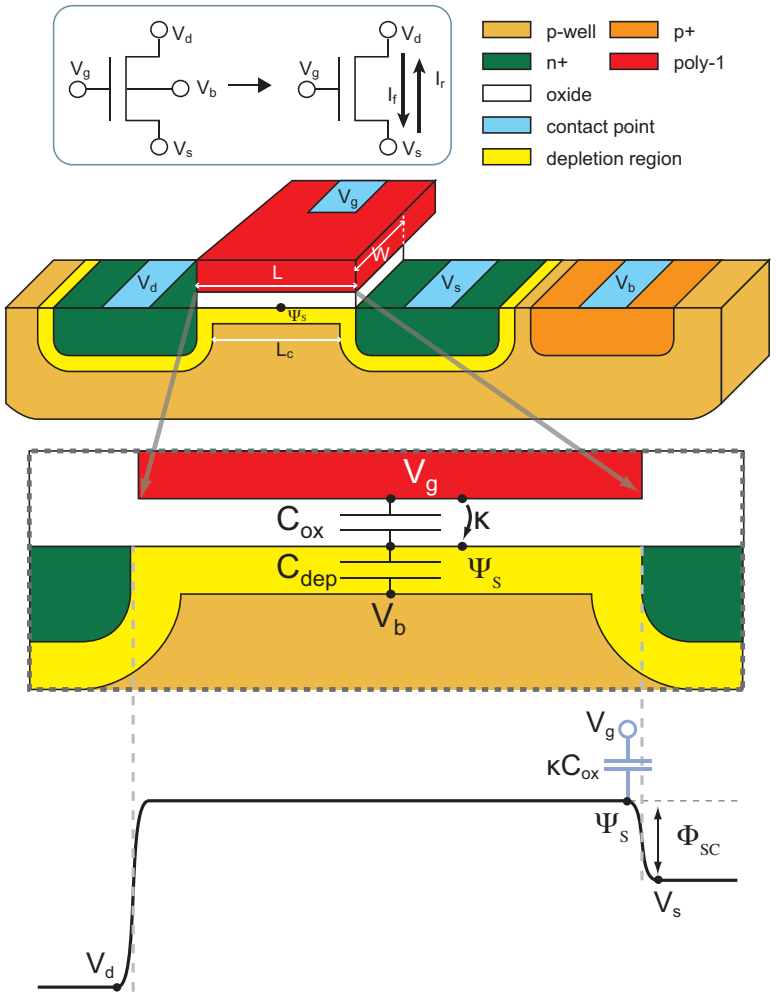


Figure 27: The illustration represents the drawn nFET layout, terminal voltages, and the effect of these voltages on the surface potential, ψ_s. Assuming that both C_{ox} and C_{dep} are fixed, the coupling from the gate voltage to the surface potential can be described as $\psi_s = \kappa V_g$. The difference between the channel potential, Ψ_s, and source voltage, V_s, results in Φ_{SC} that sets the device operation in subthreshold.

where V_g is the gate voltage, V_s is the source potential, V_d is the drain potential and V_b is the bulk potential of the four-terminal MOSFET device. The corresponding voltages can be visualized through layout and a band diagram that is illustrated in Figure 27. Other terms that are fixed during fabrication are the mobility, μ, the oxide capacitance per unit area, C_{ox}, that is partially defined by the drawn width W and drawn length L, the thermal voltage U_T, and a divider term κ. The threshold current, (81), is combined from (80), resulting in

$$I_{th} = \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{\kappa}, \tag{81}$$

$$I_{f,r} = I_{th} \ln^2 \left(1 + e^{\left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_{s,d}\right)/2U_T}\right), \tag{82}$$

and the I_{th} term has temperature dependence that is proportional to $T^{\frac{1}{2}}$, which is discussed more in Section 5.2.2. In subthreshold, (82) becomes

$$I_{subvt} = I_{th} \left(e^{(\kappa V_g - \kappa V_{T0} - V_s)/U_T} - e^{(\kappa V_g - \kappa V_{T0} - V_d)/U_T}\right). \tag{83}$$

If I assume saturation, where the reverse current is negligible, (83) then becomes

$$I_{sat} = I_{th} e^{(\kappa V_g - \kappa V_{T0} - V_s)/U_T} = I_{th} e^{(\Phi_{SC} + offset)/U_T}, \tag{84}$$

where I_{th} is the current at threshold, which is nearly in the desired form of (78). The current is now in a form of the surface potential, Φ_{SC}, as a function of V_{T0}, which has a linear dependence with U_T, plus a constant due to the offset at the flatband condition. V_{T0} is discussed in detail in Section 5.2.2, and is shown as as (109). If I assume that this offset term is fixed for all devices, the term can be simplified to only Φ_{SC}. The implication of (84) is that one expects V_s to follow V_g because Φ_{SC} must remain fixed for a constant current source. The form in (84) is the same as form in (78), resulting in

$$I_{sat} = I_{th} f\left(\frac{\Phi_{SC}}{U_T}\right), \tag{85}$$

where the function term is an exponential.

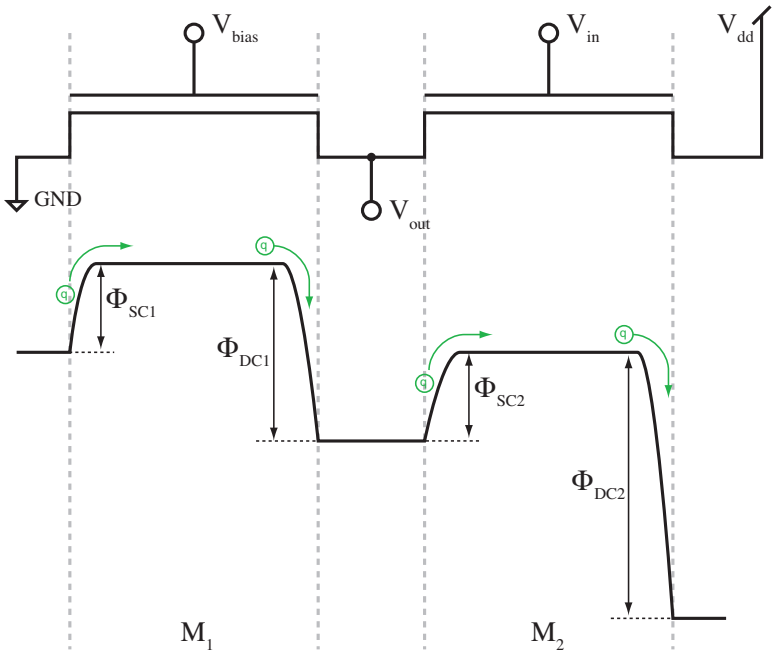


Figure 28: Assuming that no current is lost through junctions or on the output node, the flux through the circuit is fixed; therefore, the source-channel potential of both M1 and M2 must be identical.

5.1.2 Source Follower

In order to explore the interesting implications of a fixed source-channel potential, Φ_{SC}, as described by (84), consider the source-follower. In a series of subthreshold-biased devices, the charge that flows through the system is bound completely by the channel potential of the device at its lowest potential. Consider the common-drain amplifier that is shown Figure 28. The amplifier is made from two identical nFETs with the source of the bias transistor, M1, to ground. The source-channel potential of both devices (Φ_{SC1},Φ_{SC2}) must be the same for a constant current; therefore, whatever current is set through M1 by V_{bias} will also be present in M2, assuming no current is lost through V_{out}. As the gate of M2 moves, the source voltage of M2 must also move to satisfy the condition of no current loss. This is the same behavior that

one would expect for a source follower, which is

$$I_{th}e^{\Phi_{SC1}/U_T} = I_{th}e^{\Phi_{SC2}/U_T}, \tag{86}$$

$$I_{th}e^{(\kappa_1 V_{bias}-0)/U_T} = I_{th}e^{(\kappa_2 V_{in}-V_{out})/U_T}, \tag{87}$$

$$\kappa_1 V_{bias} = \kappa_2 V_{in} - V_{out}. \tag{88}$$

If one assumes that the κ terms are matched, the gain of the amplifier is κ and the offset of κV_{bias} is a constant, a simplification can be made from (88) resulting in

$$V_{out} = \kappa V_{in} - \kappa V_{bias}. \tag{89}$$

(89) can then be modified to show the relative change from input to output for a fixed bias as

$$\Delta V_{out} = \kappa \Delta V_{in}. \tag{90}$$

The form of the above equations have the interesting property of being independent of the drain voltage due to saturation. The equation in (90) is also independent of temperature due to the lack of a thermal voltage term, U_T. Furthermore, the drain-channel potential of M1, Φ_{DC1}, must change with a change in V_{in}; however, for the source follower no effects were noted from this change in drain voltage as long as M1 remained in saturation. The result is what can be described as a gain set by κ, and a capacitive relationship between the input and output through the divider as described in (90). Figure 29 illustrates this capacitive relationship. The source will follow the gate at a gain of κ because Φ_{SC} is fixed due to the bias transistor. The mechanical analog to this behavior is that of a beam with a fixed pivot at an end and the force applied at the other with the surface potential at some point along the beam. The capacitive divider represents the distances from this point with the pivot being the bulk tie.

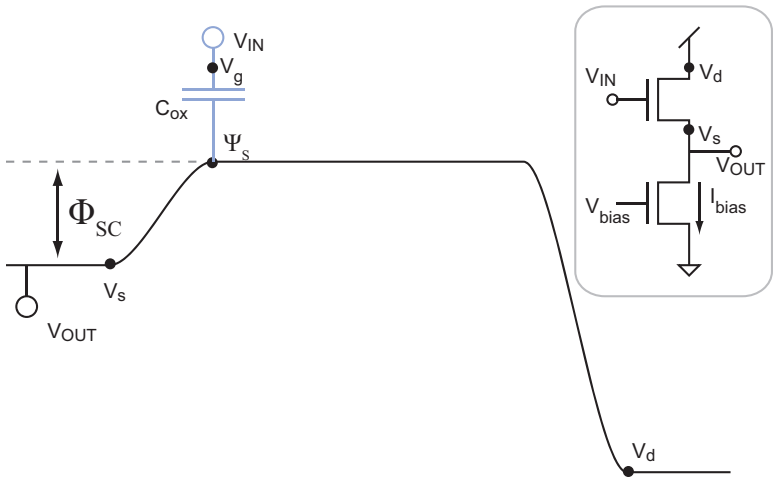


Figure 29: If the current is fixed through a saturated transistor, as in the case of the source follower, the source voltage must track the gate voltage; therefore, $\Phi_{SC} = \kappa V_g - V_s$. The depletion capacitor is not drawn because it is lumped into κ.

5.1.3 Common Source Amplifier

The common source amplifier is an interesting circuit to consider when using a capacitive approach to describing electrostatic behavior because the source is tied to the bulk. Consider the common source amplifier in Figure 30 that is operating in saturated-subthreshold. If the bias current is fixed through V_{bias}, the source to channel potential must also be fixed for both devices, so the source of the follower FET cannot follow the gate for a fixed current. Conceptually, the result can be described as a change in the divider that occurs and pulls on the drain. The capacitive divider has not changed, it behaves as if it has changed due to the behavior at the drain edge. In Figure 30, the source (V_s) is fixed because it is tied to ground. Assuming that I_{bias} is fixed and forcing a current through the amplifier, for a fixed current, the potential from the source-to-channel, Φ_{SC}, will also be fixed. This means that the surface potential, ψ_s, is also effectively fixed. This fact has the interesting implication that one can treat the coupling to the channel as a divider between the gate voltage and drain voltage. The gate voltage couples to the surface potential of the channel through the explicit oxide capacitor, and the drain voltage couples to the channel

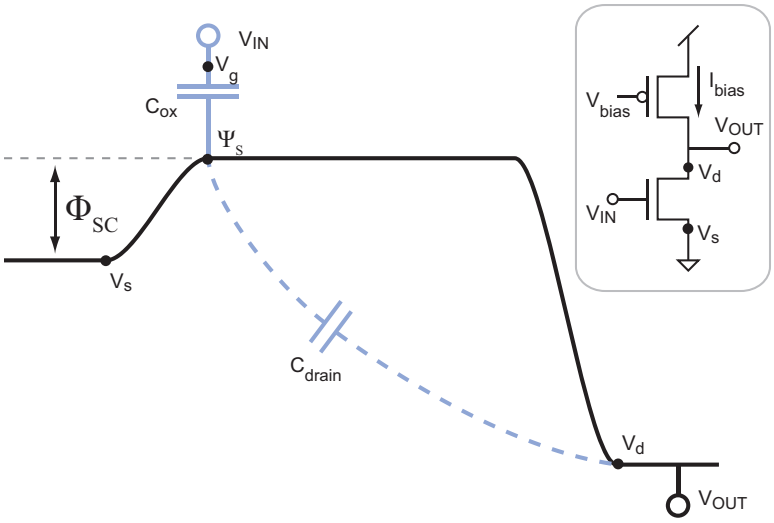
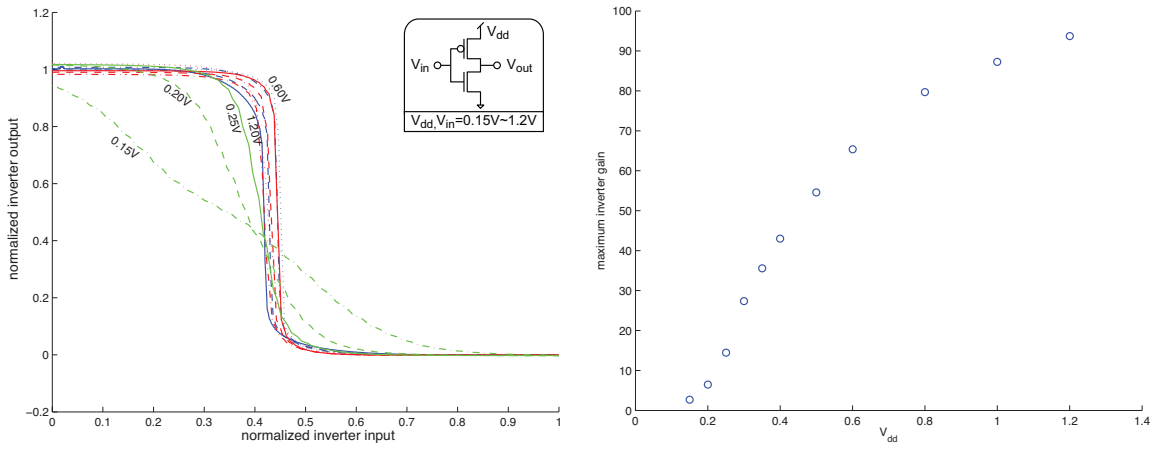


Figure 30: The band diagram for the nFET of a common source amplifier is shown. For a fixed current, the drain voltage must move opposite the gate voltage. To satisfy the condition in the illustration, the source-channel potential must be static for saturated operation. This behavior is much like a balance of forces where the gate and drain voltage must satisfy the condition set for Φ_{SC} by the bias FET.



(a) (b)

Figure 31: The inverter is a familiar circuit that is a good example for considering the behavior of circuits in subthreshold and some of the practical effects of voltage. The data presented in (a) was taken from an inverter across a range of supply voltages, and then normalized. The inverter was made of discrete, well-matched FETs with the dimensions of 2 µm square devices on a commercially available 130nm process. (b) reports the maximum gain for the inverter in (a).

surface through an implicit drain capacitor with only a fraction of the gate or drain voltages coupling back to the surface. Starting from (83), the fixed current can be described as

$$I_{fixed} = I_{th}\left(e^{\frac{\Phi_{SC}}{2U_T}}\right), \tag{91}$$

where Φ_{SC} includes the drain dependence of σV_d. In the simplified-function form of (78), (91) becomes

$$I_{fixed} = I_{th} f\left(\frac{\Phi_{SC}}{2U_T}\right). \tag{92}$$

Assuming that the charge is conserved between the FETs in the amplifier, the equation in the simple form for a pFET and nFET in series becomes

$$I_{thn} f\left(\frac{\kappa_n\left(V_{in}\right) + \sigma_n V_{out}}{2U_T}\right) = I_{thp} f\left(\frac{\kappa_p\left(V_{dd} - V_{bias}\right) + \sigma_p\left(V_{dd} - V_{out}\right)}{2U_T}\right), \tag{93}$$

where I_{thn} and I_{thp} are the threshold currents for the nFET and pFET. In the simplest case, $I_{thn} = I_{thp}$ and then (93) can be simplified to

$$\kappa_n V_{in} + \sigma_n V_{out} = \kappa_p\left(V_{dd} - V_{bias}\right) + \sigma_p\left(V_{dd} - V_{out}\right). \tag{94}$$

If assumptions are made for matched device behavior, and defining $\sigma = \sigma_n // \sigma_p$, (94) can be written as

$$\kappa \Delta V_g = -\sigma \Delta V_d + V_{const}, \tag{95}$$

where V_{const} is the fixed bias term, $\kappa_p\left(V_{dd} - V_{bias}\right)$. Substituting $V_d = V_{out}$ and $V_g = V_{in}$ for the amplifier circuit, results in the expression of

$$\Delta V_{out} = -\frac{\kappa}{\sigma}\Delta V_{in}, \tag{96}$$

where the gain term is κ/σ. The mechanical analog for this behavior is that of a beam balanced by the forces of κV_g and σV_d with the pivot set by an opposing force of Φ_{SC}. Figure 31 shows a special case of the common-source amplifier, the inverter. The

data presented in Figure 31(a) was taken by measuring the input-output relationship of the inverter for varying values of V_{dd}. Figure 31(b) reports the maximum gain for a V_{dd}. For the case of a low V_{dd}, where the devices are not in saturation, the σ goes to zero. Starting from (94), the resulting gain equation is $\Delta V_{out} = -\kappa \Delta V_{in}$, and this result is seen in Figure 31(a) for V_{dd} of 0.15V. The increase in gain is almost linear with the increase in V_{dd} until one of the devices leaves subthreshold, as seen in Figure 31(b). To explain this change in gain, one needs to consider that the Early effect manifests itself as channel current dependence on drain voltage. Because I assume that Φ_{SC} is constant, I assumed that the Early effect was unchanging. If I assume that the Early effect is an incremental change, I can model this in the same way as Hasler, resulting in

$$V_A = I_{ds} / \frac{\Delta I_{ds}}{\Delta V_{ds}} = \frac{1}{\sigma}, \tag{97}$$

which shows that the change of V_A is approximately linear with the device drain voltage[23]. This agrees with other work that showed for a step junction the drain dependence on V_A is approximately linear [11, 23, 13]. The small signal model resistance with (97), then becomes

$$\Delta r_o = \frac{V_A}{I_{ds}} = \frac{\Delta I_{ds}}{\Delta V_{ds}}. \tag{98}$$

The transconductance is

$$g_m = \frac{\Delta I_{ds}}{\Delta V_{in}}, \tag{99}$$

and the effects of this change in transconductance can be seen in Figure 31(a) at the beginning of the transition. The transconductance can now be used to solve for gain

$$A_v = \frac{\Delta V_{in}}{\Delta V_{ds}}. \tag{100}$$

This relationship is the cause of the gain changing with an increase in V_{dd} for the inverter because the saturated devices show a greater range in drain-source voltage.

Table 3: Temperature dependent terms of diffusion-based current and approximate behavior over temperature.

term	name	unit	deviation per
q	electron charge	eV	negligible
W	drawn width	meters	negligible
L	drawn length	meters	negligible
k	Boltzmann constant	eV/K	negligible
μ	mobility	$m^2/(\text{eV s})$	lattice
N_0	carrier density	cm^{-3}	fixed
ϕ_0	barrier	eV	negligible

5.2 Temperature Dependence in Subthreshold Circuits

The subthreshold region of operation is generally not considered to be temperature robust due to the explicit thermal voltage term in the exponent as in (83); however, input to output voltage relationships for circuits, such as in (90) and (96), have no explicit temperature term. The remaining temperature dependence is due to implicit terms that are weakly temperature related, such as band-gap.

5.2.1 Summary of Temperature Dependent Terms in Diffusion Movement

The temperature dependence of individual terms from the current equations are listed in Table 3. The diffusion current dominates in the subthreshold regime of operation, and drift current dominates in the above threshold operation regime. The channel current, I, of a MOSFET is can be described as the current at a position, x, which results in

$$I(x) = I_{diff}(x) + I_{drift}(x). \tag{101}$$

The transition between diffusion and drift currents occurs at the device threshold, a point that I will define as when the current in the channel is half drift and half diffusion. Starting with the diffusion current,

$$I_{diff}(x) = qNv_{diff}x, \tag{102}$$

where q is the charge of an electron, N is the charge density, x is the drawn channel width, and v_{diff} is diffusion velocity and μ is mobility as a function of velocity related to temperature. For a device with a width of W, (102) can be expanded by substitution to

$$I_{diff} = q\frac{kT}{q}\mu\frac{dN}{dx}W, \tag{103}$$

where $\frac{dN}{dx}$ is the gradient from the source to drain that is the difference in energy barriers with respect to the length at point x of the device as one moves across the channel, which is the physical distance between the barriers. Further expansion to include the dependence of the gradient as a function of voltages results in

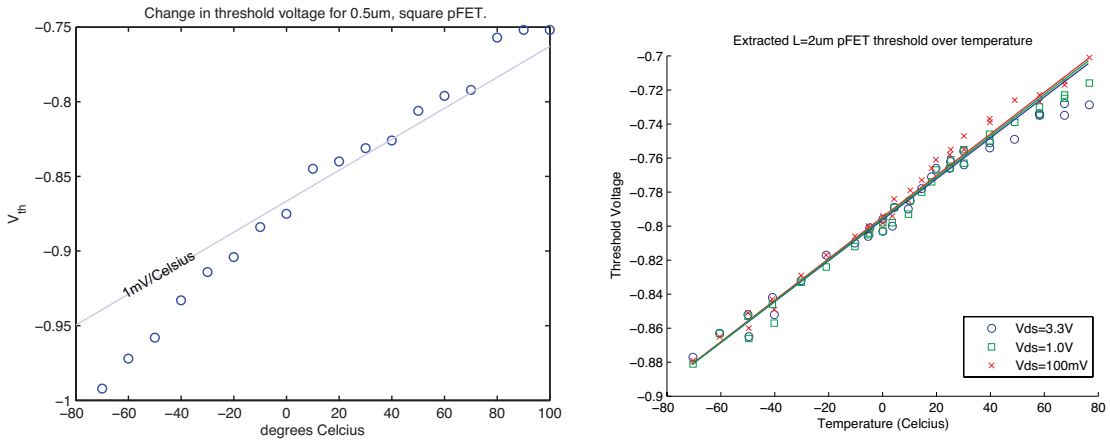
$$I_{diff} = qW\frac{kT}{q}\mu\left[\frac{1}{L}N_0 e^{-\phi_0/(kT)}e^{-(q\Psi)/(kT)}\left(e^{(qV_s)/(kT)} - e^{(qV_d)/(kT)}\right)\right], \tag{104}$$

where N_0 is the carrier density at the Fermi level, Ψ is proportional to κV_g and ϕ_0 is the built-in barrier. The temperature dependence of each of these terms is list in Table 3, and it is obvious that dominating temperature term is the thermal voltage because other terms have a very small dependence on temperature. Even with the explicit temperature terms, equating the currents through two devices that operate at the same temperature in subthreshold, such as in a source follower, will mathematically cancel out many of the terms.

5.2.2 Temperature Dependent Terms

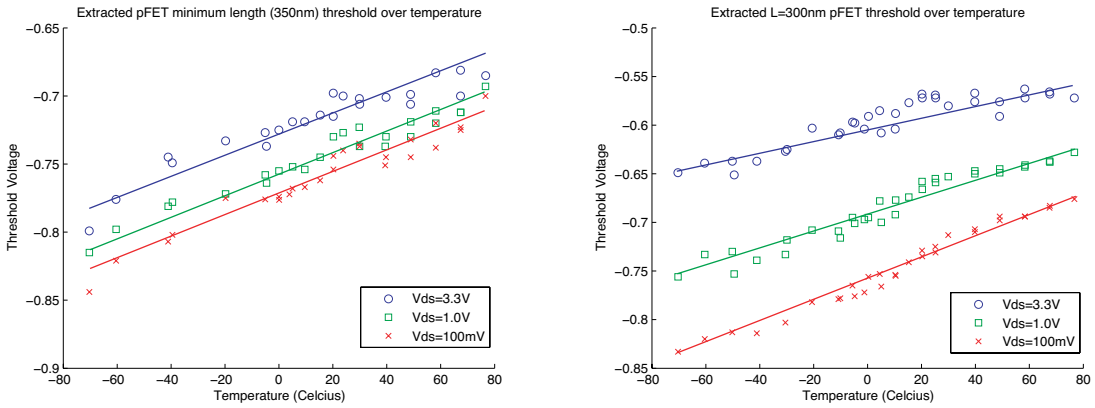
Upon inspection of the Compact EKV question, (80), the thermal voltage, U_T, is the only term that explicitly has a temperature dependence; however, the threshold voltage will also shift linearly over temperature, as shown experimentally in Figure 32, at a rate of approximately 1mV per degree Celsius. This section will result in a description for V_{T0} in the form of

$$V_{T0} = A_1 + U_T A_2, \tag{105}$$



(a) (b)

Figure 32: (a) and (b) show the threshold change over temperature for two processes. (a) reports the extracted threshold voltage over temperature for a pFET of 2 µm in width by 2 µm in length on a 0.5 µm process. The 2 µm FET shows a shift in threshold that is approximately linear with temperature. Furthermore, this threshold shift is almost unaffected by the drain voltage. (b) reports the extracted threshold voltages for pFET of 2 µm in width by 2 µm in length on a 350 nm process. The 2 µm FET shows a shift in threshold that is approximately linear with temperature and that change is almost independent of temperature. The change in threshold for (b) is approximately 1mV per degree Celsius. The change in threshold in (a) is slightly more than expected, and the change is most likely due to a constant offset due to the ESD diodes.



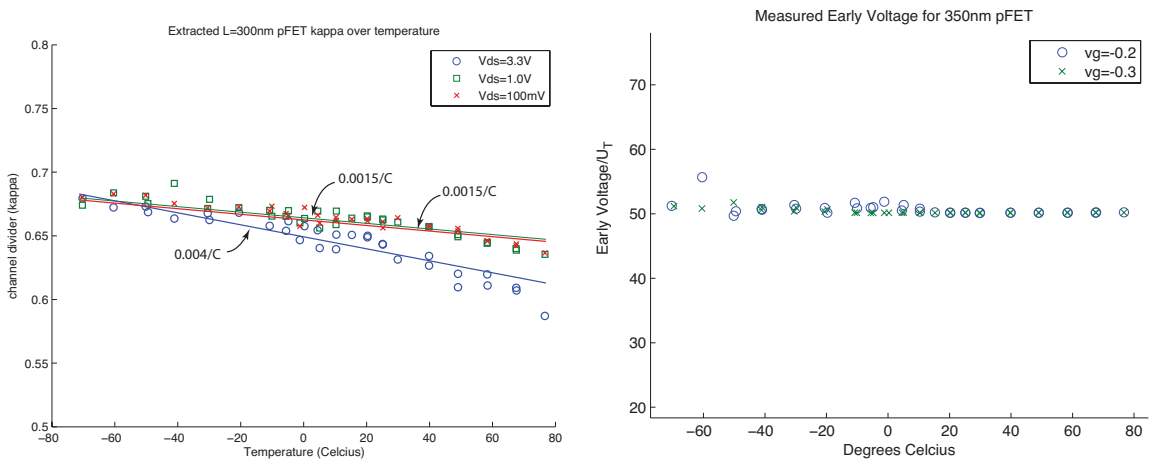
(a) (b)

Figure 33: (a) shows the extracted threshold voltage over temperature for pFET of 2 µm in width by 350nm in length on a 350 nm process. The 350nm FET shows an almost linear shift in extracted threshold voltage. (b) shows the extracted threshold voltage over temperature for pFET of 2 µm in width by 300nm in length on a 350 nm process. The 300nm FET shows an almost linear shift in extracted threshold voltage. The rate of change over temperature changes with temperature dependent behavior of depletion encroachment that changes the effective channel length.

where V_{T0} is proportional to U_T, and fermi level offsets are absorbed by the "A" terms. Referring back to (80), the term with the strongest temperature dependence is the thermal voltage, U_T. The explicit temperature term in U_T is the cause of the shift in threshold current over temperature. In order to determine how temperature affects the threshold voltage, channel surface must be described for a gate voltage. The threshold is generally defined as where the relatively abrupt inversion of the channel occurs at $\psi_s = 2\phi_f$ when the surface potential (ψ_s) is twice the Fermi potential [17]; however, a capacitive divider exists from the gate voltage to the surface potential in subthreshold. Thus, the surface potential, with respect to the gate, is

$$\psi_s = \kappa V_g. \tag{106}$$

For the pure MOS capacitor, the channel behavior is governed by the surface potential, ψ_s, as seen at the surface from the gate, V_g, that is a variable capacitive divider, κ,



(a) (b)

Figure 34: (a) and (b) show the channel divider, κ, and extracted intercept from the drain sweep for a less-than-minimum sized pFET at 300 nm on a 350 nm process. The extracted κ for larger devices showed an immeasurable change with drain voltage, and the same shift over temperature. The device in (a) is presented because the shift due to charge sharing from the drain voltage can clearly be seen.

between the channel and the gate oxide,

$$\kappa = \frac{C_{ox}}{C_{ox} + C_{dep}}, \tag{107}$$

where C_{ox} is the oxide capacitance per unit area and C_{dep} is the depletion capacitance per unit area. The gate oxide capacitor is

$$C_{ox} = \frac{WL\epsilon_{ox}}{t_{ox}}, \tag{108}$$

where W is the drawn width, L is the drawn length, t_{ox} is the oxide thickness, and ϵ_{ox} is the permittivity of silicon dioxide. All of the terms of the oxide capacitor are relatively temperature robust, and can be assumed to be effectively temperature independent. The oxide capacitance is also physically independent of the channel charge. Assume for a moment that the depletion capacitor, C_{dep}, is also fixed. This assumption results in a modified description for the threshold voltage, V_{T0}. Including the flatband voltage and the channel divider κ, the equation for the threshold voltage

becomes

$$V_{T0} = V_{FB} + \frac{2\phi_f}{\kappa}, \tag{109}$$

where V_{FB} is the flat-band voltage that includes trapped charges, the Fermi voltage ϕ_f, and κ is coupling from the gate to the surface.

Returning to C_{dep}, the depletion capacitor is dependent on channel charge. Considering that surface potential governs the amount of charge in the channel, I start by considering the depletion effect on the channel. A constant charge per unit volume is generally assumed and is $\rho = qN_D$. To satisfy this condition, I must assume that the substrate is uniformly doped, and the mobile charge is at the channel surface. Defining a coordinate system where x is perpendicular to the channel surface with $x = 0$ at the depletion edge and $x = x_0$ at the channel surface, the field in the depletion layer is $\frac{\rho x}{\epsilon_s}$. Integrating from the bulk to the surface results in

$$\psi_s = -\frac{1}{2}\frac{qN_D x_0^2}{\epsilon_s}, \tag{110}$$

where ϵ_s is permittivity of the silicon, q is the charge of the electron and N_D is donors per unit volume [36]. All of the terms of the depletion capacitor are relatively temperature robust, and can be assumed to be effectively temperature independent. Therefore, without considering changes due to higher order effects, the channel divider, κ, is temperature robust as ψ_s approaches $2\phi_f$.Returning to (109), the Fermi level is defined as

$$\phi_f = U_T \ln\left[\frac{N_D}{\sqrt{N_C N_V}} e^{V_{BG}/(2U_T)}\right], \tag{111}$$

where U_T is the thermal voltage, V_{BG} is the band-gap, and N_C and N_V are the density of states for the conduction and valence bands respectively. Expanding (111), the resulting equation is

$$V_{T0} = V_{FB} + \frac{V_{BG}}{\kappa} + \frac{2U_T}{\kappa} \ln\left[\frac{N_D}{\sqrt{N_C N_V}}\right], \tag{112}$$

where the temperature dependence of N_C and N_V is proportional to $T^{\frac{3}{2}}$; however, the change in N_C and N_V over temperature is small when compared to U_T. This implies that the shift of V_{T0} should be approximately linear with temperature due to the U_T term and that the threshold voltage increases with temperature. This is the same as the measured device behavior that is shown in Figure 33(b). Equation (112) identified that the threshold voltage will increase relatively linearly with temperature; however, the conditions of the source and drain voltages were ignored. The effects of Drain-Induced Barrier Lowering (DIBL) cause charge to be injected into the channel to show an increase in current in both subthreshold and above-threshold modes of operation. The assumption of abrupt depletion boundaries make the C_{dep} term idealized; however, longer channel devices have less charge sharing and depletion length changes so that κ is effectively linear with respect to the source and drain condition [45]. The realities of depletion encroachment in short-channel devices can be seen in Figure 33(a). Both of the devices in Figure 33 show a linear relationship to extracted threshold and temperature; however, the shorter than minimum length device has an offset in the threshold due to depletion encroachment because of the drain voltage. The effect of the drain voltage causing depletion encroachment is also seen in κ, as shown in Figure 34. The 2μm length device shows none of the drain coupling effects and has a threshold shift that is the same regardless of drain voltage within measurement and extraction error. In reality, a shift in threshold and κ must exist due to depletion encroachment, but it is immeasurable as it is less than the precision of the test equipment. The threshold voltage of the longer device is further from the bulk-reference because the charge sharing at the source and drain edges have a smaller effect on the channel length.

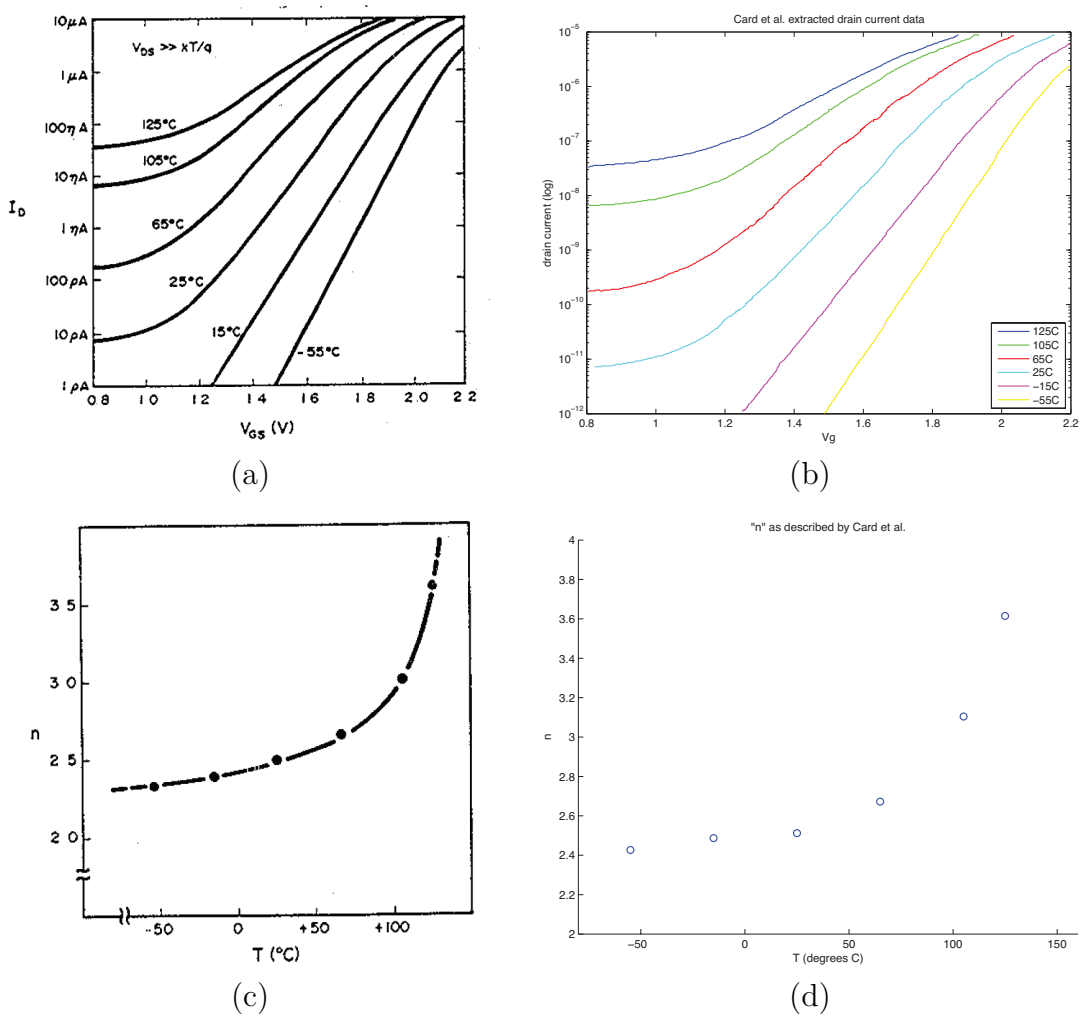


Figure 35: The data from Card et al. [6] was extracted from the photocopy using the Datathief software. (a) is the original gatesweep data and (b) is the same data imported into MATLAB and plotted. (c) is the original "n" plot from the photocopy and (d) is a recreation of (c) that was extracted from imported data in (b).

5.2.3 Temperature Independence of κ

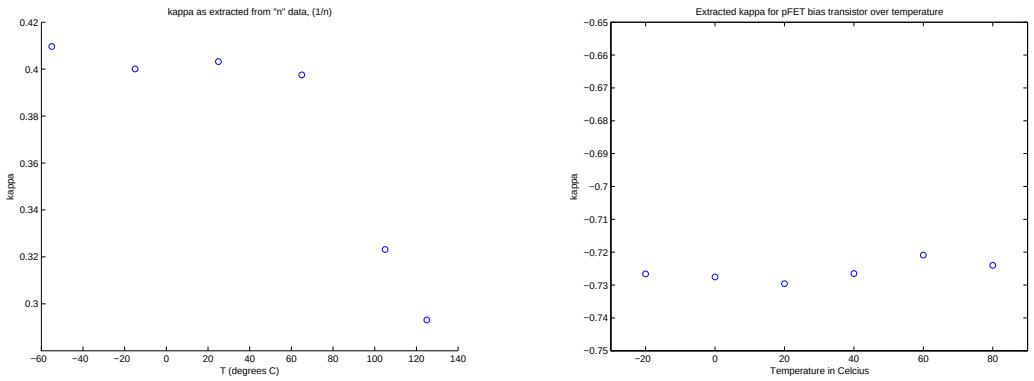
The channel divider, κ, as defined in (107), has a mixed history of temperature insensitivity [6]. The 1979 work by Card et al. reported a temperature dependence of the subthreshold slope, which was greater than expected than by the thermal voltage alone. This behavior was reported as "interface state" temperature dependence; however, in a 350nm process, I did not notice such a dependence. In order to explore this discrepancy, I extracted the data from a photocopy of the article using the Datathief software and imported the data into MATLAB for processing [53, 41]. Figure 35 shows the extracted data along with the photocopy source. These data curves represent a gatesweep of a 1979 MOSFET with largely unknown characteristics, and unknown protection circuitry. I show that some of what was reported by Card et al. is probably leakage due to protection circuity. I believe that ESD protection was used because of a history of concern regarding static discharge and MOSFET devices. I fabricated devices that did not include diode protection, which are used for the comparison. I show that the effect reported as temperature dependent "interface state" changes over temperature that correlate to an increase in current is no longer present in modern processes or is too small to extract from my measured data in an after-the-fact manner, or at the temperatures that I was able to achieve.

Figure 35(a) is the photocopied drain-current graph that was extracted and recreated as Figure 35(b). The data from Figure 35(a) was then used to calculate Figure 35(d) in the same manner as Figure 35(c). One should note that a period and negative sign are missing from copy process in Figure 35(a).

Instead of using κ notation as in (107), Card defined the channel divider as

$$n = \frac{C_o + C_d + C_{ss}}{C_o}, \tag{113}$$

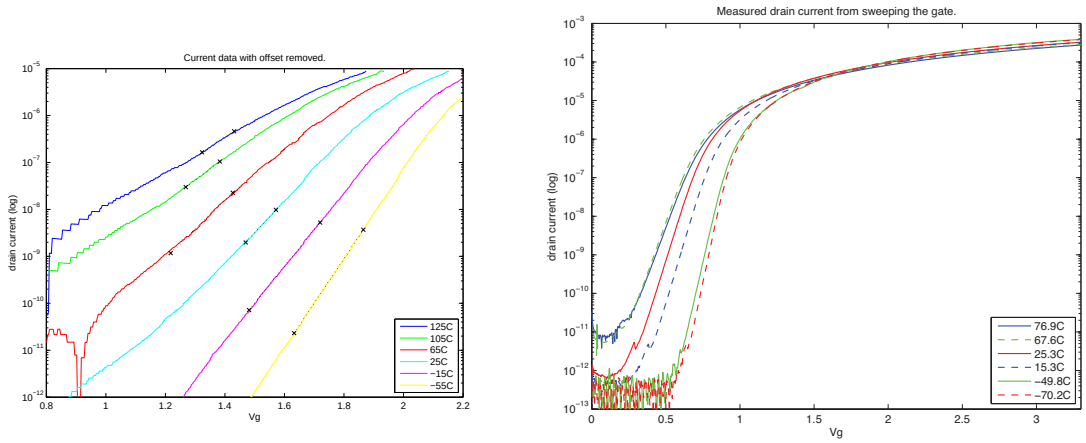
where n is just the slope of $ln(I_D)$ vs. V_s, which makes κ equal to 1/n. In 1979, Card found a temperature dependence on κ to be greater than what was expected by the



(a) (b)

Figure 36: The change in κ over temperature reported by Card et al. [6] in 1979 is presented in (a). The change in κ of a pFET device that was used for a bias generator over temperature is presented in (b). This pFET device data is presented because it was the most carefully characterized device over temperature, and it showed a sharp contrast from the data presented by Card. et al.

thermal voltage alone; however, as of 2009, I had not seen such a direct dependence. In Figure 36, I have taken Card's data to produce a κ plot from their nFET in (a) and plot of κ over temperature from a well characterized pFET in (b). The data in 36(b) is from a 2µm square pFET, which is the same pFET that I use as a current source, which was fabricated on a 350nm process in 2009. Figure 36 shows that there is a difference in measured behavior between 1979 in (a), and 2009 in (b). The question is why does the temperature dependence on κ shown in (a) is not shown in (b). Figure 37 shows the extracted drain current graph from Card, (a) with an analogous drain current graph in (b) from a nFET device that measured 350nm long by 2µm wide on a 350nm process. When compared to (a), the graph in (b) shows a device that has a significantly larger subthreshold region; however, it still does show a current floor at temperature. The current floor in (a) looks very much like data from devices that I previously fabricated that had diode protection on the source and drain terminals. The data shown in (b) is from devices which have no explicitly drawn ESD diodes; however, an implicit diode does exist from the drain to the bulk. I believe that this



(a) (b)

Figure 37: The drain current data reported by Card et al. [6] that has the current floor offset removed is presented in (a). The drain current data over temperature from a nFET device that measured 350nm long by 2um wide on a 350nm process is presented in (b). Notice that the device in (b) has a significantly large subthreshold range, even across temperature.

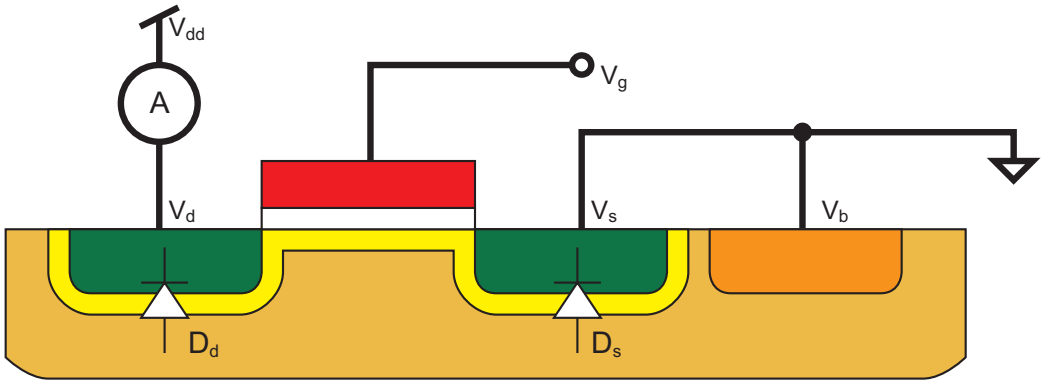


Figure 38: An illustration of the ammeter location and connections used to create the nFET gatesweeps over temperature. The junction diodes are noted from both the drain and source to the bulk.

implicit diode from the drain junction to the bulk is the cause of the measurement floor over temperature, and this current can be subtracted out. An illustration of the nFET test setup and the location of the implicit diodes is shown in Figure 38. The ammeter attempts to pin the drain voltage, V_d, to V_{dd}, which was 3.3 volts. The junction drain junction diode, D_d, is reverse biased and should have the current increase with temperature. The equation for a reverse biased diode is transcendental in form, so does not have a close-form solution for current change over temperature; however, we would expect that current change would have a form of

$$I_s \propto e^{-E_G/(2n_qU_T)}, \tag{114}$$

where E_G is the silicon band gap, n_q is the diode "quality" and U_T is the thermal voltage. The value for n_q is expected to be between 1 and 2. (114) is the slope of an extrapolated line between the measured minimum drain currents at each temperature for a diode. The slope of the plotted diode curve then can be extracted as

$$\frac{E_G}{2n_qU_{T_0}}\frac{\Delta T}{T_0}, \tag{115}$$

and used to extract n_q. The data for minimum currents plotted against temperature change is presented in Figure 39. The result of modeling the minimum current revealed that leakage current consistent with a diode exists for the 350nm device, which is shown in Figure 39 (b). The diode in (b) has a quality factor of 1.598, which is consistent with a value that is expected for a diode. The Card data revealed something that is more complicated than just a diode alone because the quality factor is 0.794, for the data shown in Figure 39 (a). The change in current is still logarithmic with temperature, but with a substantially higher slope than just a single diode. It is possible that the data was from a CD4007 MOSFET pair, which was available at that time, and has very similar drain current characteristics[25]. According to the CD4007 data sheet, the ESD protection for the CD4007 consists of diode clamps and then a resistor to another set of diode clamps. If this structure exists in the transistors that

Card used, it is possible that this structure is causing a voltage change over temperature. This would result with the gate voltage being different from what is measured at the input pin and thereby affecting the measured current, and this structure would also affect the drain and source terminals to cause slight voltage changes to affect the current.

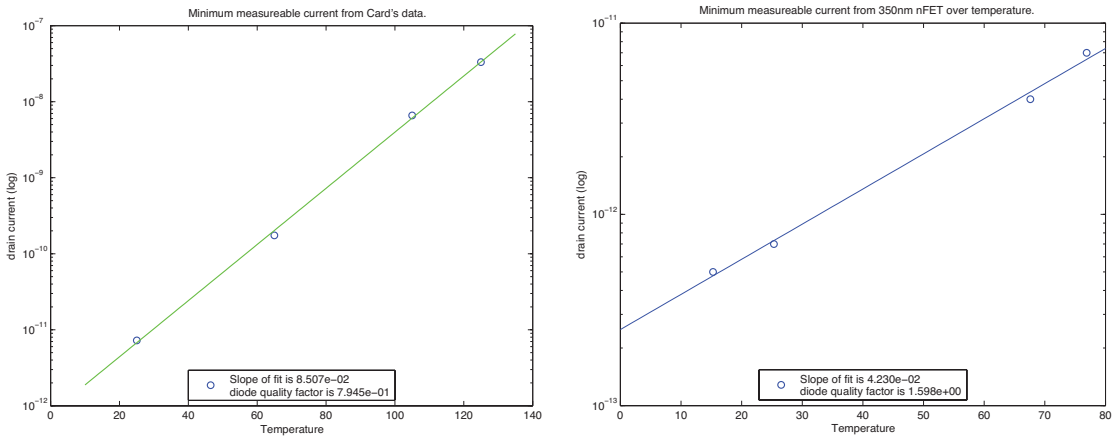
Using the leakage information from Figure 39, new κ values over temperature were extracted with the result shown in Figure 40. Even with the diode leakage extracted, κ did not revert to a relatively fixed value. I expect for a large device that κ would be effectively unchanging, but a short device will have different κ values for changes in drain voltage. In this case, the drain voltage is fixed. Interestingly, both graphs in Figure 40 show the same trend, with the κ value decreasing suddenly for temperatures greater than 60 degrees Celsius. This could be from either the interface state change that Card first reported, or depletion encroachment on the short devices due to temperature. Although I cannot completely dismiss the effect of interface states, these states have significantly less impact on this 350nm process, where the highest measured κ deviation was less than 2%.

5.2.4 Temperature Independence of σ and gain in subthreshold

The gain in subthreshold was derived in (96) and the equation has no explicit temperature terms, containing only κ and σ. This equation can be rewritten as

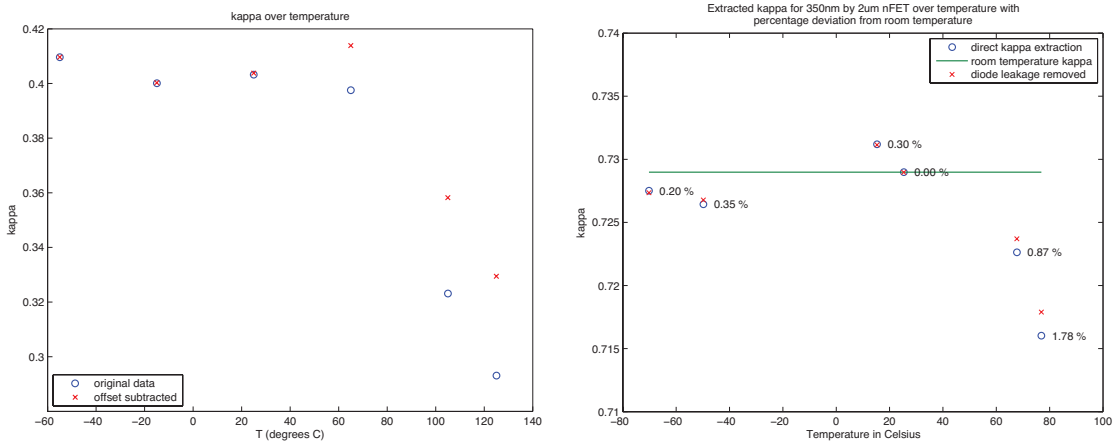
$$\frac{\Delta V_{out}}{\Delta V_{in}} = gain = -\frac{\kappa}{\sigma}. \tag{116}$$

There are two significant terms, κ and σ. κ was previously shown to be robust against changes in temperature in Figure 36; however, a slight deviation due to depletion encroachment was shown in Figure 34(a) for a 300nm FET over temperature. A square pFET of 2µm in length and width showed no such change in κ over temperature. The extracted κ value is presented in Figure 36(b).



(a) (b)

Figure 39: (a) is minimum measurable current over temperature and line fit from Card et al. [6]. (b) is the minimum measurable current over temperature from a nFET device that measured 350nm long by 2um wide on a 350nm process.

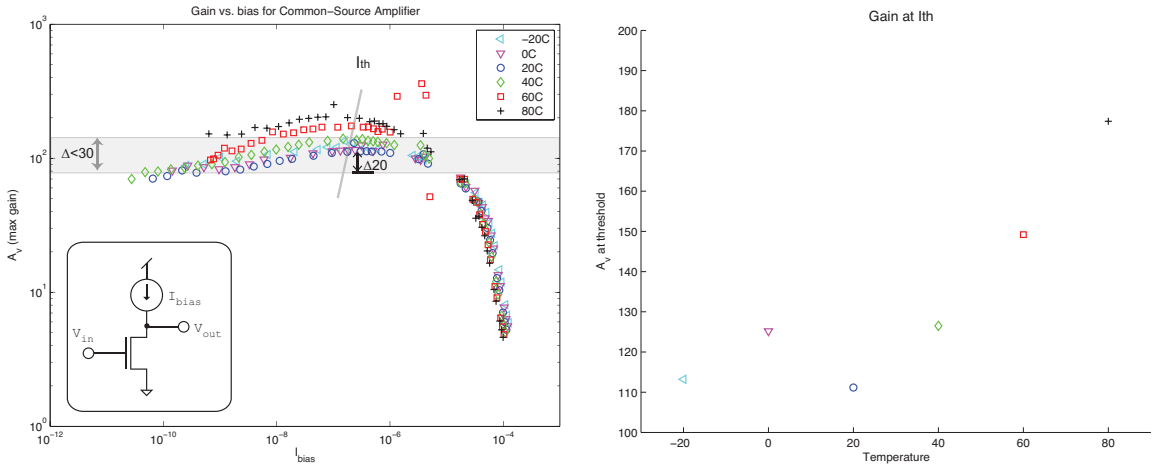


(a) (b)

Figure 40: The extracted κ after subtracting possible diode leakage currents for data from Card in (a) and from a nFET on a 350nm process in (b).

A common source amplifier was used to create an inverting amplifier. A characterized pFET was used to set the bias current, I_{bias}, over temperature. The gain of the amplifier was then measured over temperature in order to analyze the behavior of σ with the results reported in Figure 41. In subthreshold, (116) suggests that the gain is fixed for all values of I_{bias}; furthermore, it should be fixed for all temperatures. Figure 41(a) shows the gain for a circuit schematic in the inset of the illustration. The gain is relatively flat in subthreshold, and then falls off in the above threshold range; however, the gain still remains high until 3μA. The slight increase in gain with current is unexpected, as one would expect the gain to decrease or stay flat; however, it is possible that this increase is due to the diode effects reported in Section 5.2.3. Even with this unexpected rise, the change in gain is approximately 20 across all bias ranges for a temperature, and the change across all temperatures is only 30 with the average gain at about 100. The relative shape of this gain curve applies across all temperatures. Figure 41(b) reports the maximum gain that is at threshold as extracted from Figure 41(a). The current ranged from 260nA to 320nA at -20°C and 80°C respectively. The figure shows that from 20°C, the gain increases approximately by one for every one degree Celsius.

The higher gains seen with temperature are most likely due to increased currents in the bias transistor from diode leakage. The bias transistor current was calculated from a measured current for a specific gate voltage. Because the κ and σ have both been shown to be robust to temperature change through separate measurements, the increase in the gain is most likely due to a leakage path from the control gate of the bias transistor that is injecting current through some other current source. Alternatively, if the bias transistor was actually shorter than expected, the gain could be seen increasing because of a σ change due to depletion encroachment. I believe that the devices used for this experiment are both square devices that measured 2μm by 2μm. If there was a wiring error that caused the bias device to be 2μm wide by



(a) (b)

Figure 41: The gain of a common source amplifier for different values of I_{bias} over temperature is reported in (a). The gain changes slightly with bias and temperature. The change of gain with temperature is no more than a factor of two, and the change of gain with bias current is of a similar magnitude. The greyed area represents the range of gain change. For 20°C, the change in gain is only 20 across this range with an average gain of about 100, and the shape of the change holds across other temperature ranges. Note the marked decrease in gain in the above threshold region of operation. (b) reports the gain at threshold current, which corresponds to maximum gain.

350nm long, it is possible that the increase in gain that is being seen is due to a decreasing σ as the depletion region encroaches.

5.3 Common Source Temperature Dependence

Assuming that capacitive dividers are effectively constant as in Section 5.2.2, any change on the gate will be seen directly on the drain. As a thought experiment, (96) is what is expected for a common-source amplifier as it has a negative gain, which one would expect to be high. Furthermore, both κ and σ have been shown to have very little deviation with temperature under the appropriate conditions. One can expect that this common-source amplifier has a gain that is robust against temperature change. Revisiting the fact the current is constant through the nFET, the current

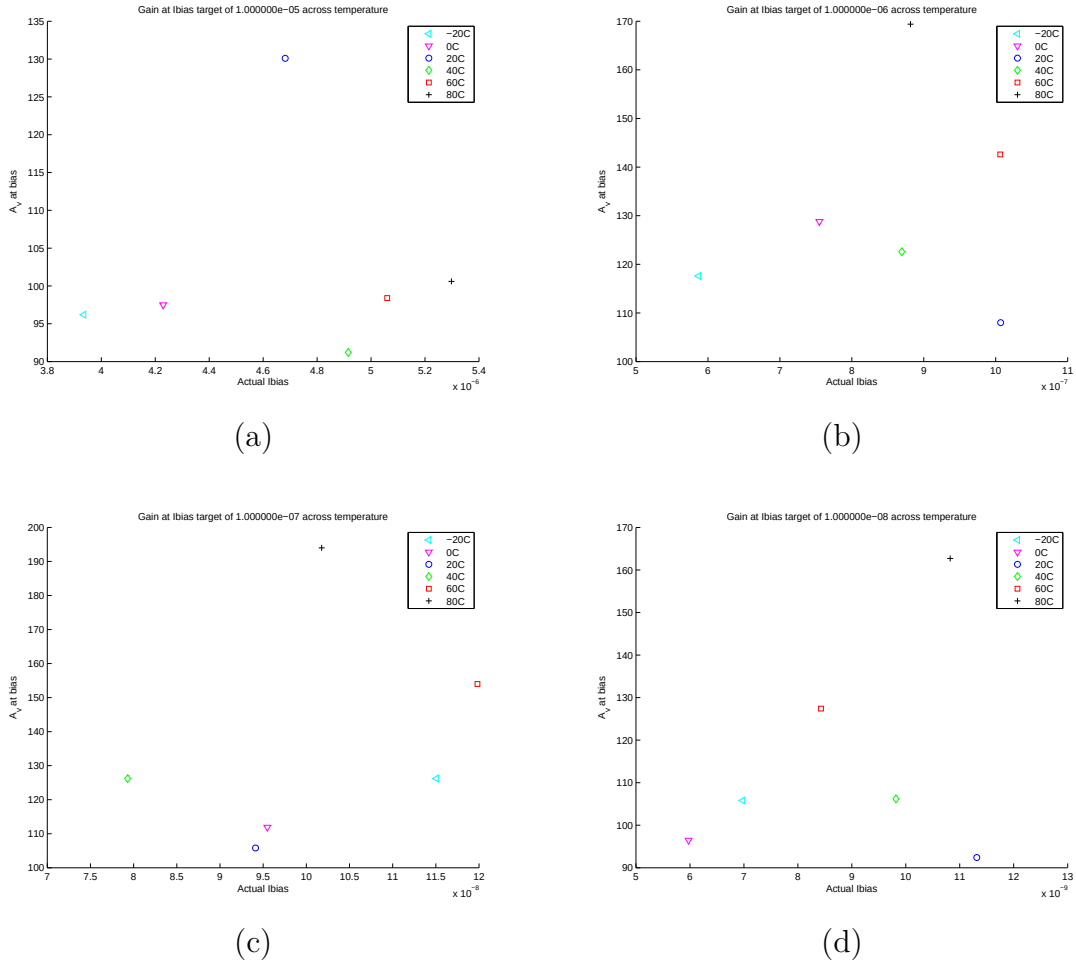
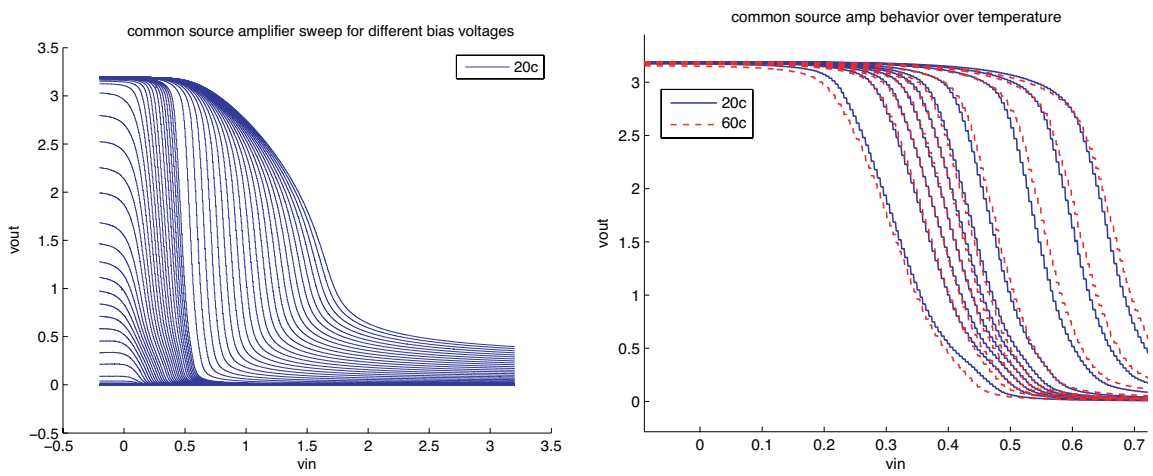


Figure 42: The gain measured for a given bias at temperature.



(a) (b)

Figure 43: The input-output behavior for a common-source amplifier fabricated in a commercially available 0.5 µm process is shown in (a), where both devices were 2µm square. (b) reports the results for the same device from the high-gain region for temperatures of 20°C and 60°C. A bias voltage exists where the behavior of the device is temperature independent.

can be approximated by

$$I_{sat} = I_{th} e^{(\phi_s - 0)/U_T} = I_{th} e^{\Phi_{SC}/U_T}, \tag{117}$$

where the thermal voltage is $U_T = \frac{kT}{q}$. The saturated current for a single device has a strong temperature dependence due to the U_T term in the exponent; however, the common-source amplifier circuit has an advantage due to the pFET device because the temperature dependent terms cancel. The equality is

$$I_{th_{nFET}} e^{\Phi_{SC_{nFET}}/U_T} = I_{th_{pFET}} e^{\Phi_{SC_{pFET}}/U_T}, \tag{118}$$

which reduces to

$$I_{th_{nFET}} e^{\Phi_{SC_{nFET}}} = I_{th_{pFET}} e^{\Phi_{SC_{pFET}}}, \tag{119}$$

$$\Phi_{SC_{nFET}} = \Phi_{SC_{pFET}} + \ln \left[\frac{I_{th_{pFET}}}{I_{th_{nFET}}} \right]. \tag{120}$$

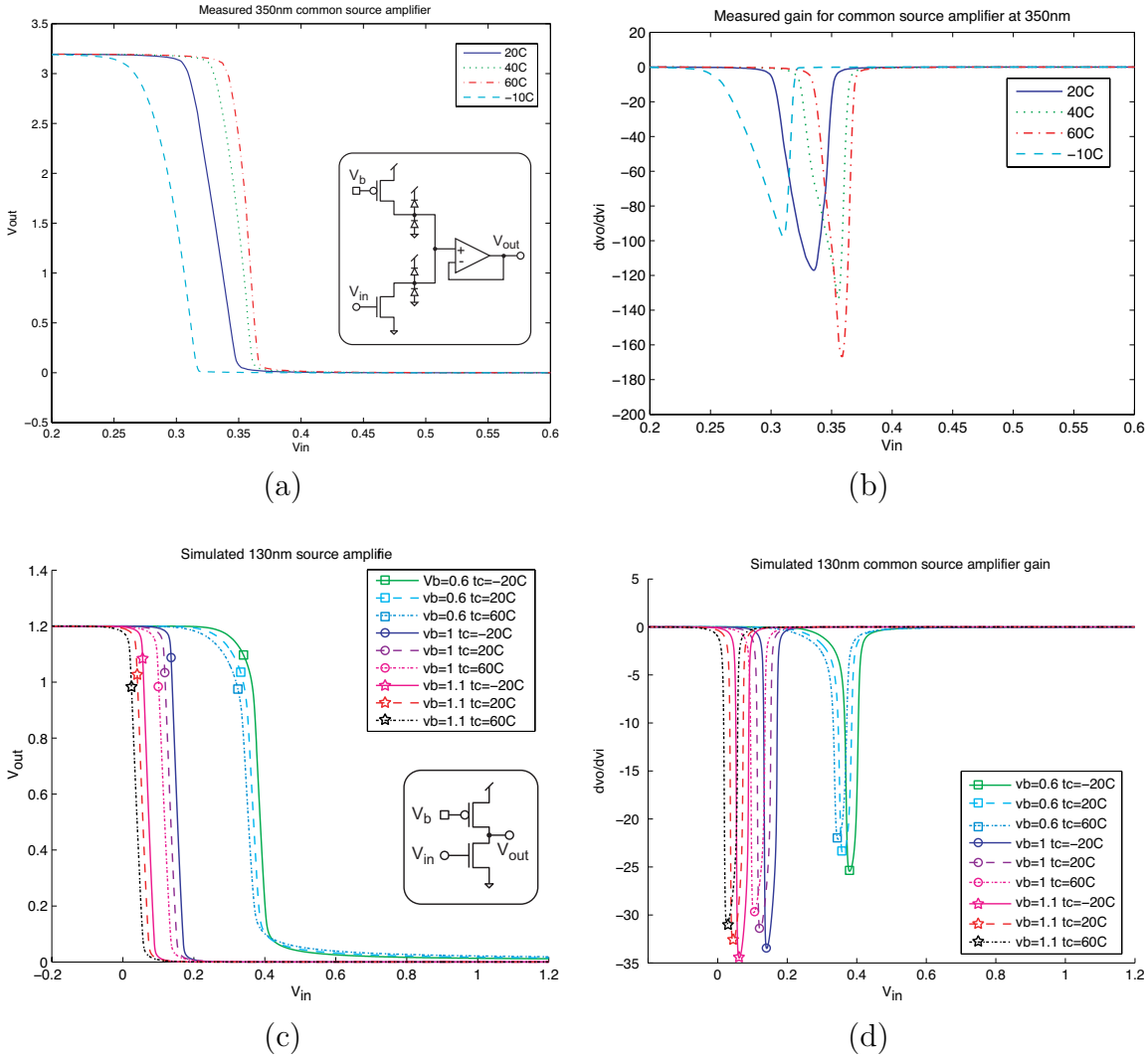


Figure 44: The gain measured from a common-source amplifier fabricated in a commercially available 350 nm process is shown in (a) with the corresponding voltage gain in (b). The devices were 2µm square for the purposes of matching. Simulated results from a commercial available 130 nm process are shown in (c) and (d) using the BSIM 4.4 models supplied with the design kit for devices 390 nm square. Simulation results for the 350 nm devices with BSIM 3.3f models yielded results that were erroneous because the subthreshold behavior is not well modeled. One can see that the general behavior between the fabricated and simulated devices is similar. The gain plot in (d) shows that the subthreshold devices not only have higher gain with temperature, they have higher gain than the above-threshold devices.

In (120), the threshold currents for each of the devices will create an offset to the current, which will then be seen as an offset where the high-gain region exists, as seen in Figure 44(c).

5.4 Discussion

The balance of forces approach to subthreshold circuit design works well for describing the behavior of the source follower and common-source amplifier, even if the description is not completely true to the electrostatics. The form of $f_1 = f_2$ has the other advantage of canceling the explicit temperature terms that results in a form that is temperature robust. The circuits in application show very little temperature variance for large devices, where a large device is any device that has a long channel.

The validation of the $f_1 = f_2$ was done through common-source amplifiers that showed that both κ and σ are fixed when normalized for temperature, or when in a form where the temperature terms cancels out. The common-source amplifiers that are the basis for this work were fabricated in commercially available 0.5μm and 350nm processes and simulated at the 130nm node using commercial models. The 0.5μm common source amplifier that is presented in Figure 43 is temperature robust for a set of bias voltages near threshold as predicted by (112). For the 350nm process, the measured threshold voltages over temperature for different drain conditions of devices 2 μm wide and 300 nm, 350 nm and 2μm in length are shown in Figure 33 (a), (b), and (c) for devices fabricated on a commercial 350 nm process. The minimum size pFET and long pFET in Figure 33 (b) and (c) both showed a threshold shift of approximately 1mV per degree Celsius. This linear behavior is what one would expect from the threshold description of

$$V_{T0} = V_{FB} + \frac{V_{BG}}{\kappa} + \frac{2U_T}{\kappa} \ln \left[\frac{N_D}{\sqrt{N_C N_V}} \right], \tag{121}$$

which is the same as (112) with the explicit temperature dependence of the thermal voltage. The shorter than minimum size pFET in Figure 33 (a) shows a threshold

change that is approximately 1mV per degree Celsius at $V_{ds} = 100mV$, but a different shift for larger drain voltages. This shift is worth discussion because the threshold is extracted by fitting the subthreshold region data and the threshold is picked when the channel current is half diffusion movement. The equation (121) would point to κ changes as the cause of this shift, but there maybe something in the drain condition contrary to the assumptions of Fjeldly [13]. Figure 34(b) showed that the Early voltage for a minimize-sized device did not change when normalized for temperature, which suggests that the change in channel length over temperature should be negligible and should not affect κ while in subthreshold. The minimum-sized and longer devices showed no shift in κ that was measurable over temperature, showing that charge-sharing from the depletion region at the channel edges has no significant effect on the channel length, and that the depletion capacitor also does not produce a significant change in capacitance due to temperature, as expected from (110) with the dependencies listed in Table 3. The short-channel device showed a change in κ where the divider strength decreased with temperature which is most likely due to depletion encroachment. One alternative explanation for a shift in threshold on the shorter channel devices not following the same trend as longer devices is that there is an assumption that field lines are all terminated. Consider a device with fixed Φ_{SC}, the device will have a fixed depletion width at the drain due to the fixed channel potential, which is the assumption in [13]; however, it may be true that the relative doping of the $2\phi_f$ in the threshold term is causing a shift due to the ions available to terminate the field lines in a two dimensional space because the depth of the depletion from the drain to the bulk is a different width than that of the drain to channel.

The common source amplifier in Figure 44 showed a change in gain and voltage offset with temperature. The voltage offset is most likely due to the difference in shift in threshold current due to changes in the relative doping term between the

pFET and nFET in (120). The change in gain is slight, and may be due to the 1-D assumption for the drain condition that was discussed in the previous paragraph.

CHAPTER VI

CONCLUSION

This work has encompassed subthreshold design and applications with a focus on the floating-gate transistor. This exploration has included modeling, floating-gate transistor design and applications, as well as predictions for future device behavior. The novel contributions of this work are as follows:

- *Floating-Gate Inverters*: Single and dual floating-gate inverter structures were design and fabricated to show programmable control of offset and gain.
- *Capacitively Coupled CMOS*: A modification to the floating-gate inverter that allows for biases to be programmed by simply coupling into the the gate through a second control gate without a nFET as a quasi-reset (as in the work by Shibata et al.)
- *Non-programmed Floating-Gate Transistors*: The trapped charge on floating-gate transistors as seen back from fab were characterized to show that non-uniform charge was stored on the floating gate. These devices could have future use as crude current biases.
- *Passgate Modeling*: A model for passgate transistors based upon an effective mobility was developed that allows for zeroth order analysis of crossbar networks.
- *Passgates without a Gate-Capacitor*: A cross-bar switch matrix for floating-gate arrays that used above-threshold injection and lacks an explicit gate capacitor was designed and tested.

- *Estimation of Velocity Boundary*: Estimates of the boundary of velocity saturation for the above threshold MOSFET were found to be at 25nm. At lengths smaller, devices have entered velocity saturation as soon as they enter above threshold.

- *Force Balance Model*: The behavior of subthreshold devices was described and a conjecture was made that structures could be constructed that had behaviors robust to changes in temperature. I was shown that source-followers and common-source amplifiers mathematically and physically are robust to temperatures.

The possibility of a future, formal design methodology for temperature robust subthreshold circuits is a particularly interesting prospect. This will be particularly useful for arrays for floating-gate devices used for computation because the devices will be operating in subthreshold for high-resistance values.

REFERENCES

[1] AMD. http://www.amd.com/us/products/Pages/products.aspx.

[2] Bisdounis, L. and Koufopavlou, O., "Short-circuit energy dissipation modeling for submicrometer CMOS gates," *IEEE Transactions on Circuits and Systems-1: Fundamental Theory and Applications*, vol. 47, no. 9, 2000.

[3] Bowman, K., Tschanz, J., Wilkerson, C., Lu, S., Karnik, T., De, V., and Borkar, S., "Circuit techniques for dynamic variation tolerance," in *Proceedings of the 46th Annual Design Automation Conference*, pp. 4–7, ACM, 2009.

[4] Brian, F., Asano, S., Peter, H., Gilles, G., Kim, R., Le, T., Liu, P., and Jens, L., "Microarchitecture and implementation of the synergistic processor in 65-nm and 90-nm SOI," *IBM Journal of Research and Development*, vol. 51, no. 5, pp. 529–543, 2007.

[5] Calhoun, B., Wang, A., and Chandrakasan, A., "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid State Circuits*, vol. 40, no. 9, pp. 1778–1786, 2005.

[6] Card, H. and Ulmer, R., "On the temperature dependence of subthreshold currents in mos electron inversion layers," *Solid-State Electronics*, vol. 22, no. 5, pp. 463–465, 1979.

[7] Chandrakasan, A., Sheng, S., and Brodersen, R., "Low-power CMOS digital design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, 1992.

[8] Degnan, B., Wunderlich, R., and Hasler, P., "Passgate resistance estimation based on the compact EKV model and effective mobility," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pp. 2765–2768, IEEE, 2009.

[9] Delagi, G., "Harnessing technology to advance the next-generation mobile user-experience," in *2010 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 18–24, 2010.

[10] Duffy, C. and Hasler, P., "Modeling Hot-Electron Injection in pFET's," *Journal of Computational Electronics*, vol. 2, no. 2, pp. 317–322, 2003.

[11] Early, J., "Effects of space-charge layer widening in junction transistors," *Proceedings of the IRE*, vol. 40, no. 11, pp. 1401–1406, 1952.

[12] Enz, C., Krummenacher, F., and Vittoz, E., "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83–114, 1995.

[13] Fjeldly, T. and Shur, M., "Threshold voltage modeling and the subthreshold regime of operationof short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, pp. 137–145, 1993.

[14] Godfrey, M., "Cmos device modeling for subthreshold circuits," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 39, no. 8, pp. 532–539, 1992.

[15] Gray, J. D., Twigg, C. M., Abramson, D. N., and Hasler, P., "Characteristics and programming of floating-gate pfet switches in an fpaa crossbar network.," in *ISCAS (1)*, pp. 468–471, 2005.

[16] Gray, P., Hurst, P., Meyer, R., and Lewis, S., *Analysis and design of analog integrated circuits*. Wiley, 2001.

[17] Grove, A., *Physics and technology of semiconductor devices*. Wiley New York, 1967.

[18] Hall, T. S., Twigg, C. M., Hasler, P., and Anderson, D. V., "Developing large-scale field-programmable analog arrays.," in *IPDPS*, 2004.

[19] Hanson, S., Zhai, B., Bernstein, K., Blaauw, D., Bryant, A., Chang, L., Das, K., Haensch, W., Nowak, E., and Sylvester, D., "Ultralow-voltage, minimum-energy CMOS," *IBM journal of research and development*, vol. 50, no. 4.5, pp. 469–490, 2010.

[20] Hasler, P., Andreou, A., Diorio, C., Minch, B., and Mead, C., "Impact ionization and hot-electron injection derived consistently from Boltzmann transport," *VLSI Design*, vol. 8, no. 1-4, pp. 455–461, 1998.

[21] Hasler, P., Basu, A., and Kozil, S., "Above Threshold pFET InjectionModeling intended for Programming Floating-Gate Systems," in *IEEE International Symposium on Circuits and Systems, 2007. ISCAS 2007*, pp. 1557–1560, 2007.

[22] Hasler, P. and Dugger, J., "Correlation learning rule in floating-gate pfet synapses," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 48, pp. 65–73, January 2001.

[23] Hasler, P., *Foundations of learning in analog VLSI*. PhD thesis, California Institute of Technology, 1997.

[24] Instruments, T. http://power.ti.com.

[25] Instruments, T., "Cd4007ub." original document from Harris Semiconductor.

[26] Intel. http://ark.intel.com/Default.aspx.

[27] ITRS, "International technology roadmap for semiconductors," 2009.

[28] Johnson, S., "Residual charge on the faulty floating gate mos transistor," in *Proceedings of the IEEE International Test Conference on TEST: The Next 25 Years*, (Washington, DC, USA), pp. 555–561, IEEE Computer Society, 1994.

[29] Kerns, D. A., Tanner, J. E., Sivilotti, M. A., and Luo, J., "Cmos uv-writable non-volatile analog storage," in *Proceedings of the 1991 University of California/Santa Cruz conference on Advanced research in VLSI*, (Cambridge, MA, USA), pp. 245–261, MIT Press, 1991.

[30] Kinget, P. R., "Device mismatch and tradeoffs in the design of analog circuits," *Solid-State Circuits, IEEE Journal of*, vol. 40, pp. 1212–1224, June 2005.

[31] Kogge, P., Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hill, K., and others, "Exascale computing study: Technology challenges in achieving exascale systems," *DARPA Information Processing Techniques Office, Washington, DC*, p. 278, 2008.

[32] Koomey, J., Berard, S., Sanchez, M., and Wong, H., "Assessing trends in the electrical efficiency of computation over time," *IEEE Annals of the history of computing*, 2009.

[33] Lenzlinger, M. and Snow, E., "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," *Journal of Applied Physics*, vol. 40, pp. 278–283, January 1969.

[34] Lide, D., *CRC handbook of chemistry and physics*. CRC press, 1993.

[35] Liu, S., *Analog Vlsi: Circuits and Principles*. Bradford Books, 2002.

[36] Maher, M., *A charge-controlled model for MOS transistors*. PhD thesis, California Institute of Technology, 1989.

[37] Manohar, R. and Tierno, J., "Asynchronous parallel prefix computation," *IEEE Transactions on Computers*, vol. 47, no. 11, pp. 1244–1252, 1998.

[38] Marr, B., Degnan, B., Hasler, P., and Anderson, D., "An asynchronously embedded datapath for performance acceleration and energy efficiency," in *IEEE International Symposium on Circuits and Systems, 2009. ISCAS 2009*, pp. 3046–3049, 2009.

[39] Marr, B., Degnan, B., Hasler, P., and Anderson, D., "Considerations for Zeroth-Order Current Approximations for Power-Constrained Computing," *IEEE Transactions on VLSI*, submitted.

[40] Martin, A., "Asynchronous datapaths and the design of an asynchronous adder," *Formal Methods in System Design*, vol. 1, no. 1, pp. 117–137, 1992.

[41] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[42] Mead, C. and Analog, V., *Neural Systems*. Addison-Wesley, 1989.

[43] Moore, G., "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 19 April 1965.

[44] MOSIS. http://www.mosis.com.

[45] Odame, K., McDonald, E., and Minch, B., "Highly linear, wide-dynamic-range multiple-input translinear element networks," in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, 2003.

[46] Patterson, D. and Hennessy, J., *Computer Architecture: A Quantitative Approach, 3rd edition*. Morgan Kaufmann, 2003.

[47] Rodriguez-Villegas, E. and Barnes, H., "Solution to trapped charge in fgmos transistors," *Electronics Letters*, vol. 39, pp. 1416 – 1417, September 2003.

[48] Singh, M. and Nowick, S., "The design of high-performance dynamic asynchronous pipelines: high-capacity style," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 11, pp. 1270–1283, 2007.

[49] Smith, P., Graham, D., and Hasler, P., "A kappa projection algorithm (kpa) for programming to femtoampere currents in standard cmos floating-gate elements," *Analog Integrated Circuits and Signal Processing*, vol. 56, pp. 83–91, 2008. 10.1007/s10470-007-9089-x.

[50] StJohn, I. and Fox, R., "Leakage effects in metal-connected floating-gate circuits," *Circuits and Systems II: Express Briefs, IEEE Transactions on [see also Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on]*, vol. 53, no. 7, pp. 577–579, 2006.

[51] Sutherland, I., Sproull, R., and Harris, D., *Logical effort: designing fast CMOS circuits*. Morgan Kaufmann, 1999.

[52] T. Shibata, H. Kosaka, H. I. and Ohmi, T., "A functional mos transistor featuring gate-level weighted sum and threshold operations," *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 8, pp. 913 – 922, 1995.

[53] Tummers, B., "Datathief III." http://datathief.org/, 2006.

[54] Uyemura, J. P., *CMOS Logic Circuit Design*. Norwell, MA: Kluwer Academic Publishers, 1999.

[55] van Langevelde, R. and Klaassen, F., "Effect of gate-field dependent mobility degradation on distortion analysis in MOSFETs," *Electron Devices, IEEE Transactions on*, vol. 44, no. 11, pp. 2044–2052, 1997.

[56] Veendrick, H., "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 4, 1984.

[57] Venkatesh, G., Sampson, J., Goulding, N., Garcia, S., Bryksin, V., Lugo-Martinez, J., Swanson, S., and Taylor, M., "Conservation cores: reducing the energy of mature computations," *ACM SIGPLAN Notices*, vol. 45, no. 3, pp. 205–218, 2010.

[58] Watt, J. and Plummer, J., "Universal Mobility-Field Curves for Electrons and Holes in MOS Inversion Layers," in *VLSI Technology, 1987. Digest of Technical Papers. Symposium on*, pp. 81–82, 1987.

[59] Weste, N. and Harris, D., *CMOS VLSI design: a circuits and systems perspective*. Addison-Wesley, 2005.

[60] Wittenbrink, C., Kilgariff, E., and Prabhu, A., "Fermi gf100 graphics processing unit (gpu)," in *Hot Chips 22*, 2010.

[61] Wunderlich, R., "CMOS gate delay, power measurements and characterization with logical effort and logical power," Master's thesis, Georgia Institute of Technology, 2009.

[62] Zamdmer, N., Ray, A., Plouchart, J., Wagner, L., Fong, N., Jenkins, K., Jin, W., Smeys, P., Yang, I., Shahidi, G., and others, "A 0.13-μm SOI CMOS technology for low-power digital and RF applications," in *VLSI Technology, 2001. Digest of Technical Papers. 2001 Symposium on*, pp. 85–86, IEEE, 2002.

Temperature Robust Programmable Subthreshold Circuits through a Balanced Force Approach

Brian P. Degnan

97 Pages

Directed by Professor Jennifer Hasler

The subthreshold region of operation has simple physics which allows for a balanced-force approach to behavioral modeling that has shown to be robust to temperature, and a model that encapsulates MOSFET behavior across all operational regions has been developed. The subthreshold region of operation also allows for injection of charge onto floating nodes that allows for persistent storage that can be used in a variety of applications. The combination of charge storage and device modeling has allowed for the development of programmable circuits for digital applications.