Purdue University Purdue e-Pubs

ECE Technical Reports

Electrical and Computer Engineering

12-1-1997

A model for leakage control by MOS transistlor stacking

Mark C. Johnson Purdue University School of Electrical and Computer Engineering

Dinesh Somasekhar Purdue University School of Electrical and Computer Engineering

Kaushik Roy Purdue University School of Electrical and Computer Engineering

Follow this and additional works at: http://docs.lib.purdue.edu/ecetr

Johnson, Mark C.; Somasekhar, Dinesh; and Roy, Kaushik, "A model for leakage control by MOS transistlor stacking" (1997). *ECE Technical Reports*. Paper 79. http://docs.lib.purdue.edu/ecetr/79

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

A MODEL FOR LEAKAGE CONTROL BY MOS TRANSISTOR STACKING

Mark C. Johnson Dinesh Somasekhar Kaushik Roy

TR-ECE 97-12 December 1997



School of Electrical and Computer Engineering Purdue University West Lafayette, Indiana 47907-1285

A model for leakage control by MOS transistor stacking

Mark C. Johnson, Dinesh Somasekhar, and Kaushik Roy School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana, 47907-1285, USA Phone: (765)494-3372, (765)494-3372, (765)494-2361 Email: {mcjohnso. somasekh. kaushik@ecn.purdue.edu}

^{&#}x27;This research was supported in part by ARPA (F33615-95-C-1625), NSF CAREER award (9501869-MIP), IBM, Rockwell, AT&T/Lucent, and ASSERT program (DAAH04-96-1-0222).

Abstract

Prevailing CMOS design practice has been very conservative with regard to choice of transistor threshold voltage, so as to avoid the difficult problems of threshold variations and high leakage currents. It is becoming necessary to scale threshold voltages more aggressively in order to obtain further power reduction, performance improvement, and integration density. Substantial leakage reduction can be achieved in single Vt designs by stacking low Vt transistors. We have derived a simplified theoretical model which predicts the quiescent leakage current and the worst case time required to settle to quiescent levels in a single stack of transistors. This model can be used in a design environment to make quick estimation of leakage with respect to design changes. Model results are compared to circuit simulation. Leakage current predictions were found to match simulation results very closely for a wide random selection of design parameter values and temperatures. Transistor stacks with multiple transistors turned off were found to have anywhere from 2 to 30 times lower leakage current than stack with only one transistor turned off. The time required for a transistor stack to settle to quiescent current levels varied from a few microseconds up to tens of milliseconds.

Contents

| 1 | Introduction | | |
|---|---|---|----|
| | 1.1 | Sources of leakage | 1 |
| | 1.2 | Simple example of leakage behavior | 2 |
| 2 | Effect of stack height on quiescent leakage | | |
| | 2.1 | Theoretical Model | 5 |
| | 2.2 | Sensitivity to process and other variations | 8 |
| | 2.3 | Simulation and theoretical model results | 9 |
| | 2.4 | Sensitivity of model to transistor characterization | 10 |
| 3 | Lea | kage transients | 12 |
| | 3.1 | Theoretical model | 13 |
| | 3.2 | Simulation and theoretical model results | 17 |
| | 3.3 | Sensitivity to process and other parameters | 19 |
| | 3.4 | Energy cost associated with leakage transients | 19 |
| | 3 .5 | Exploiting the stacking effect | 20 |
| 4 | Con | clusions | 21 |

List of Figures

| 1 | Simple NAND gate | 2 |
|---|--|----|
| 2 | Leakage behavior of pull down network in SAND gate | 4 |
| 3 | Schematic and notation for stacking effect analysis | 5 |
| 4 | Correlation of simulated and estimated leakage | 11 |
| 5 | Correlation of simulated and estimated leakage savings | 12 |
| 6 | Transistors and capacitances affecting internal node i | 15 |
| ī | Discharge of internal node capacitances | 16 |
| 8 | Correlation of simulated and estimated settling time | 18 |

1 Introduction

An accurate estimate of standby leakage power must consider circuit topology as well as signal levels when the circuit is idle. Kawahara [5] clemonstrated this in the design of a low power decoded-drivers for a DRAM. An extra transistor was placed between the supply line and the pull-up transistor for the driver. This causes a slight reverse bias between the gate and source of the pull-up transistor when both transistors are turned off. Because subthreshold current is exponentially dependent on gate bias, a substantial current reduction was obtained. This phenomenon is referred to as the "stacking effect".

In this paper we derive a more general model of the stacking effect with respect to subthreshold current reduction and the time required to settle to quiescent current levels. This model considers the general case of transistor stacks with an arbitrary number of transistors. It takes into account both body effect and drain induced barrier lowering (DIBL). DIBL (retluction of threshold voltage as V_{DS} increases) is especially significant for sub-micron devices. The leakage of a transistor stack is shown to be directly tlependent on the magnitude of the DIBL effect.

1 Sources of leakage

In current and near future MOS technologies, the dominant component of leakage current is subthreshold current [6]. Shrinking transistor size has greatly increased subthreshold current while reducing junction diode leakage which was a dominant leakage component in earlier technologies. As dimensions continue to shrink. other causes of leakage may become significant. At present, gate induced drain leakage (GIDL) poses the greatest threat to leakage control by means of transistor stacking. GIDL is largest when V_{DS} is largest and V_{GS} is strongly reverse biased. The stacking effect relies on reverse biasing of V_{GS} to achieve leakage savings. Consequently, GIDL may become a lower bound on leakage in the future.

1.2 Simple example of leakage behavior



Figure 1: Simple NAND gate

Before presenting the leakage model in detail, let us examine a very simple case where the stacking effect becomes significant. Figure 1 depicts a simple static two input NAND gate. We would like to understand the leakage behavior of this gate for various inputs. In the case where both PMOS transistors are turned off, the leakage is simply the sum of the off currents of each PMOS device. However, the situation for series connected transistors

is more complex. Figure 2 demonstrates what happens to the internal node voltages and currents when only the bottom NMOS transistor is initially off and then the upper NMOS transistor is turned off. A logarithmic time axis is used to make it easier to compare initial and final conditions which are separated by a wide time interval. Initially, the supply and ground line leakage current's are equal to the off current of a single transistor. As soon as the gate of the top transistor is switched off, there is an immediate drop in internal node voltage due to capacitive coupling (bootstrapping). After bootstrapping, the internal node voltage is discharged only very slowly since the only discharge mechanism is the off current through the bottom transistor. Notice that while the internal node is discharging, leakage from the supply voltage line is negiligible. This is due to the strong reverse bias between the gate and source of the top transistor. Once the internal node voltage reaches its quiescent level, then the supply and ground currents reach equilibrium at a reduced quiescent current level. In the remainder of this paper, we will derive expressions which model the leakage behavior of stacks consisting of an arbitrary number of transistors. The model will predict quiescent current' and voltage levels and the worst case "settling" time required to transition to new quiescent levels after switching off one or more transistors.



Figure 2: Leakage behavior of pull down network in NAND gate

2 Effect of stack height on quiescent leakage

2.1 Theoretical Model



Figure 3: Schematic and notation for stacking effect analysis

Let Figure 3 depict a transistor stack to be analyzed. Steady state leakage values can be estimated as a function of the number of transistor; that are turned off. Details of the derivation can be found in the appendix. The general approach is to equate the subthreshold current through each transistor and then solve for the quiescent voltage (V_{DSq_i}) across each transistor. Throughout this paper, a "q" in a subscript indicates a quiescent value. These voltages can then be used estimate the magnitude of the leakage cur-

rent. The following analysis is done for an NMOS pull down stack, but is ecually applicable to a PMOS stack.

The subthreshold current of the i^{th} MOS transistor in a stack can be modeled as

$$I_{subth} = A \times e^{\frac{1}{n\nu_T}(V_G - V_S - V_{TH_0} - \gamma' \times V_S + \eta \times V_{DS})} \times (1) \times (1 - e^{\frac{-V_{DS}}{\nu_T}})$$

where $A = \mu_0 C'_{ox} \frac{W}{L_{eff}} (\frac{kT}{q})^2 e^{1.8} e^{\frac{-\Delta v_{TH}}{nv_T}}$. Equation *I* is adapted from the BSIM 2 MOS transistor model [8,3]. V_{TH_0} is the zero bias threshold voltage. ν_T is the thermal voltage $\frac{kt}{q}$. The body effect for small values of V_S is very nearly linear. It is represented by the term $\gamma' V_S$, where γ' is the linearized body effect coefficient. η is the DIBL coefficient, representing the effect of V_{DS} ($V_{DS} = V_D - V_S$) on threshold voltage. C_{ox} is the gate oxicle capacitance. μ_0 is the zero bias mobility. n is the subthreshold swing coefficient of the transistor. ΔV_{TH} accounts for variations in threshold voltage from one transistor to another. For the conditions illustrated in figure 3, all transistors are turned off with $V_G = 0$.

First we equate the currents of the first and second transistors in the stack. We obtain equation 2 by solving for V_{DS_2} in terms of V_{DD} , as described in the appendix. It is assumed here that $V_{DD} >> V_{Sq_1}$ so that we can calculate V_{DSq_2} using V_{DD} rather than V_{DSq_1} .

$$V_{DSq_2} = \frac{n\nu_T}{(1+2\eta+\gamma')} ln(\frac{A_1}{A_2}e^{\frac{\eta V_{DD}}{n\nu_T}} + 1)$$
(2)

One can similarly equate the current through the $(i-1)^{th}$ and i^{th} transistors, solving for V_{DSq_i} in terms of $V_{DSq_{i-1}}$. This results in equation 3. Equation 3 can be used iteratively to find V_{DS_i} for each transistor, starting with the third in the stack. Finally, V_{DSq_1} can be determined by subtracting the sum of V_{DSq_i} through V_{DSq_N} from V_{DD} .

$$V_{DS_i} = \frac{n\nu_T}{(1+\gamma')} ln(1 + \frac{A_{i-1}}{A_i}(1 - e^{\frac{-1}{\nu_T}V_{DSq_{i-1}}}))$$
(3)

The voltage offset at the source of each transistor is given by $V_{S_i} = \sum_{j=i+1}^{N} V_{DS_j}$. If we are only interested in the magnitude of the leakage current, we can use $V_{DS_{q_N}}$ in equation 1 to compute the leakage through the bottom transistor. To verify this computation, one could compute the leakage of other transistors in the stack.

Once we have V_{DS_i} for each transistor, the voltage offset at the source of each transistor is given by $V_{Sq_i} = \sum_{j=i+1}^{N} V_{DSq_j}$. V_{DSq_i} and V_{Sq_i} are now known for each transistor, so we can compute the steady state leakage current using equation 1. Now let us determine the leakage savings obtained by turning off multiple transistors in a stack rather than a single transistor turned off. Dividing the leakage of a single transistor by the leakage of a stack of transistors turned off, we find the savings ratio as a function of the number of transistors (N) to be:

$$S(N) = e^{\frac{1}{n\nu_T}(1+\eta+\gamma')\sum_{j=2}^N V_{DSq_i}}$$
(4)

Take note that this analysis only considers transistors that are turned off. Transistors that turned on can be treated as if they were a short circuit. Thanks to the very small currents involved (on the order of nA or smaller), the voltage drop across transistors that are turned on will he orders of magnitude smaller than the voltage drop across transistors in the subthreshold region.

2.2 Sensitivity to process and other variations

The magnitude of subthreshold current is sensitive to many parameters. but threshold voltage and temperature variation are of particular interest because the dependence is exponential or greater. Inspection of the subthreshold current equation reveals that a small relative change in other parameters (length, width, C_{OX}) will cause an equal relative change in subthreshold current. Device climensions variations can also indirectly affect leakage by influencing thresholcl voltage.

In the subthreshold current equation, one might not initially expect an exponential increase with respect to temperature since T appears as a $\frac{1}{T}$ term in the exponent. However, for typical operating temperatures (on the order of 300 or $400^{\circ} K$) the current approximately doubles for every $8^{\circ} K$ increase in temperature. This is the same as the temperature sensitivity of silicon bipolar devices.

Sensitivity with respect to threshold voltage variation (due to variations in doping and channel length) is equal to the subthreshold slope, for which current increases by a factor of 10 given a change in threshold voltage on the order of 80 to 100mV. Supply voltage only indirectly affects leakage through the DIBL effect, for which the $V_{DD}/log(Ids)$ slope can be obtained as $\frac{\eta}{subthreshold - slope}[V/decade].$

The leakage savings ratio, given in equation 4, exhibits very little sensitivity to variations in threshold voltage or dimensions dimensions, provided that the variations are uniform for all transistors in a stack. In our model, the effect of a uniform shift in threshold voltage or dimension disappears in the derivation of the predicted savings ratio. For a stack of two 3u/0.5utransistors? HSPICE simulations showed only a 5% drop in savings ratio if threshold voltage was swept from approximately 0.6V down to 0.2V.

Temperature variation has a significant effect on the leakage savings ratio, however substantial savings are still observed for a wide range of temperatures. For a stack of two 3u/0.5u transistors, equation 4 predicts that the leakage savings ratio will drop from 14.8 to 3.8 for a temperature sweep from -50 to $150^{\circ}C$. HSPICE simulation predicts a drop in savings ratio from 10.8 to 4.2, over the same temperature range.

2.3 Simulation and theoretical model results

In this section. we will compare theoretical model predictions to simulation results for steady state leakage conditions. The simulation result:; were obtained using HSPICE with the BSIM 1 model for a 0.5u MOSIS process. The available MOSIS models do not include measured subthreshold characteristics, so we have estimated the subthreshold swing and related parameters from threshold voltage parameters, using the technique derived by Kang et. al. [4]. A subthreshold slope of approximately 80mV/decade was estimated and incorporated into the 0.5u BSIM model. In order to approximate the be-

havior of low threshold high leakage devices, we modify the flat hand voltage parameter (VFBO).

Each of the following figures compare model predictions to simulation results for 64 sets of randomly selected design parameters that describe a transistor stack. The parameters that were allowed to vary were the following: temperature (-50 to $150^{\circ}C$), number of transistors in the stack turned off (2 to 4 transistors), V_{TH_0} (from approximately 0.26V to 0.56V), supply voltage (from 1.2V to 1.8V), and transistor width (from 3μ to 10μ). Each transistor in the stack was treated as having identical characteristics for purposes of validating our simplified leakage model. The horizontal axis of each graph corresponds to a range of model predictions. The vertical axis corresponds to the range of values extracted from simulation results. Each data point identifies a model prediction and the corresponding simulation result.

Figure 4 compares model predictions of steady state leakage to simulation results. Figure 5 compares model predictions of leakage saving ratios to simulation results. The savings ratio was obtained by dividing the leakage of a single transistor by the leakage of the transistor stack. In both graphs, a very close correlation is observed.

2.4 Sensitivity of model to transistor characterization

Subthreshold slope, the DIBL coefficient (η) , and the linearized body effect coefficient (γ') are by far the most critical parameters to the estimation of leakage current and leakage savings. Zero bias threshold voltage is critical to leakage estimation, but has no effect on the savings ratio unless threshold



Figure 4: Correlation of simulated and estimated leakage

variations from one transistor to the next are considered. These parameters all have an exponential influence on leakage and savings estimate:;.

Other parameters (dimensions. C_{OX} , and carrier mobilities) only have a proportional effect on leakage estimates. and no effect at all on savings estimates except for variations from one transistor to another.



Figure 5: Correlation of simulated and estimated leakage savings

3 Leakage transients

In previous sections. we have shown that leakage can he greatly reduced by stacking transistors to be turned off when a circuit is idle. The time for a circuit to reach this quiescent low leakage state can be several orders of magnitude greater than the clock period or latency of most digital logic. This delay is a result of charges trapped on internal nodes which can only charge or discharge to quiescent levels by means of leakage currents that are very srnall in comparison to normal switching currents. A long settling time is not necessarily a disadvantage to the use of transistor stacking. However, let us first examine the behavior of a transistor stack for best and worst case settling time and then consider implications of the long settling time.

3.1 Theoretical model

Consider again the transistor stack illustrated in figure 3. A realistic worst case settling time corresponds to the case where all internal nodes are initially charged to the maximum possible voltage $(V_{DD} - V_{TH})$ just before the node is completely isolated by transistors that are turned off. This maximizes the amount of charge that must be dissipated by means of leakage before the circuit settles to quicscent levels. The worst case condition can he achieved by the following sequence of events. All but the bottom transistor are initially on so that all internal nodes can charge to $V_{DD} - V_{TH}$. Now turn off the transistor next to the bottom. Because the gate of this transistor is capacitively coupled to nodes above and below (due to gate overlap capacitance), the voltage of both nodes are pulled down somewhat (referred to as "bootstrapping"). Just after bootstrapping, the voltage at the ith internal node can be estimated as

$$V_{boot_i} = \frac{-V_T C_{ov} + (V_{DD} - V_T) C'_i}{C''_i}$$
(5)

where C_{ov} is the gate-source overlap capacitance of transistor i. Internal node *i* corresponds to the source of transistor *i* and the drain of transistor *i*+1. C'_i is the value of the internal node capacitance just before bootstrapping, including the gate-drain overlap capacitance of transistor *i* + 1. C''_i is the value of the internal node capacitance just after bootstrapping, including the gate-source overlap capacitance of transistor 2 and the gate-drain overlap capacitance of transistor i + 1. Figure 6 identifies the capacitances and transistors directly affecting internal node i. Typically each internal node consists entirely of the diffusion that is shared by the source and drain of adjacent transistors. Notice that only overlap capacitance is included in the gate to diffusion coupling. One might expect that gate to channel capacitance $(C_{ox} W L)$ would produce additional coupling. However, the transistor being switched is already on the edge of cutoff $(V_{GS} = V_{TH})$. Simulation results indicate that the degree of bootstrapping is close to that indicated by overlap capacitance alone.

This analysis assumes that all transistors in the stack are being turned off. If we wished to consider a case where an internal transistor is not switched off, we must consider that transistor in determining the total node capacitances for bootstrapping and settling time calculations. Unlike the quiescent current analysis, transistors that remain on can not be ignored. When determining node capacitances, a transistor that remains turned on can he viewed as a piece of interconnect. Gate and diffusion capacitances must then be included as a part of the internal node capacitance.

Within nanoseconds after bootstrapping, the node above the transistor being switched will charge hack up to $V_{DD} - V_{TH}$. If the next transistor up is then turned off, the bootstrapping process will repeat itself. Once all transistors in the stack are off, we find all the internal nodes charged up to approximately V_{boot} , as given in equation 5.

Now if the circuit is idle for a sufficiently long time, the internal nodes will



Figure 6: Transistors and capacitances affecting internal node i

begin to discharge and eventually reach quiescent levels as illustrated for a stack of four transistors in figure 7. Initially only the node closest to ground will discharge through the bottom transistor. All of the other transistors in the stack are strongly reverse biased ($V_{GS} < 0$) and will have leakage currents that are orders of magnitude smaller than the bottom transistor. The next node in the stack will not start to discharge significantly until the bottom node has nearly reached the quiescent level given in section 2.1. 1 he third node from ground will not discharge until the second node has nearly reached its quiescent level. This process is repeated until all nodes in the stack reach quiescent levels. This is illustrated in figure 7 where the current discharge is displayed for each internal node in a stack of four transistors. Each current waveform was obtained as the difference between the channel currents of the

transistors above and below the node being discharged.



Figure 7: Discharge of internal node capacitances

We estimate the time for each node to discharge as follows. During discharge, the rate by which node voltage (V_i) drops can be determined as a function of the node voltage.

$$\frac{dV_i}{dt} = -\frac{I_{dis}(V_i)}{C_i(V_i)} \tag{6}$$

 $I_{dis}(V_i)$ is the magnitude of the discharge current as a function of node voltage. $C_i(V_i)$ represents the node capacitance formed by the shared diffusion of the transistors above and below. C_i could include interconnect capacitance if the transistor stack is not implemented in a single contiguous strip of diffusion. C_i may also include gate and diffusion capacitances of transistors which are not switched off. The inverse of equation 6, $\frac{dt}{dV_i}$, enables us to estimate the elapsed time corresponding to an incremental decrease of V_i . Integrating over the range by which the voltage drops, we find the time taken for the node voltage to discharge from V_{boot_i} clown to the quiescent voltage level, V_{q_i} . To make the integral tractable, it was necessary to assume that capacitance remains constant. Details of the derivation are deferred to the appendix. Equation 7 gives the resulting expression for the discharge time of internal node i.

$$t_{dis_{i}} = \frac{n C_{i} L_{eff}}{\mu_{0} C_{ox} W \nu_{T} e^{1.8} \eta} \times$$

$$e^{\frac{1}{n\nu_{T}} ((1+\gamma'+\eta)V_{q_{i+1}}+V_{TH_{0}})} \times$$

$$(e^{\frac{-\eta V_{q_{i}}}{n\nu_{T}}} - e^{\frac{-\eta V_{boot_{i}}}{n\nu_{T}}})$$

$$(7)$$

 ν_T is the thermal voltage $\frac{kT}{q}$. V_{boot_i} is the voltage at the internal node just after switching of the transistor above. taking into account bootstrapping. V_{q_i} is the quiescent level for the internal node voltage, as determined by the leakage model in section 2.1. C_i is the total capacitance of the internal node. Since C_i decreases with voltage. we conservatively choose $C_i = C_i(V_{q_i})$. All other terms have the same definition as given in section 2.1.

3.2 Simulation and theoretical model results

In this section, we will compare theoretical model predictions to simulation results for the settling time of leakage transients once two or more transistors in a stack have been turned off. The simulation results were obtained in the same manner as described in section 2.3. Figure 8 uses a scatter diagram to compare settling time estimates for random selections of transistor parameters and transistor stacks of various heights. The vertical axis indicates simulation measurements of the time required for supply voltage current to settle to within 10% of its quiescent level. The horizontal axis indicates settling time derived from the theoretical model. Correlation between the simulated and estimated leakage can be observed by the clustering of points along the diagonal.



Figure 8: Correlation of simulated and estimated settling time

3.3 Sensitivity to process and other parameters

Settling time varies by orders of magnitude in inverse proportion to the magnitude of the leakage current which is also subject to wide variation. Consequently, it is strongly dependent on the same parameters as discussed in section 2.2 for leakage current.

Settling time is proportional to the size of the internal node capacitance since node capacitance multiplied by voltage is what determines how much charge needs to be discharged. Consequently, it is essential to have an accurate measure of node capacitance that includes the voltage dependence of diffusion junction capacitance.

3.4 Energy cost associated with leakage transients

Circuit, level estimation of transient leakage current costs is a complex task. However, our preceding analysis offers some insight into the problem. In the worst case settling time analysis we see that very little leakage current is drawn from the supply until the node furthest from ground (in an NMOS stack) has almost completely discharged. If the next set of inputs to the circuit were to discharge the pull down network, then leakage to ground did not cost us anything. Charges on the internal nodes would be discharged to ground regardless of whether or not leakage occurred. On the other hand, if the next set of circuit inputs cause the internal nodes to be charged up again, then the energy dissipated clue to leakage is a complete loss. In general, leakage does not cost us anything if charge is being moved in the same direction as it would during the next switching event. Conversely, leakage energy is completely lost if it flows opposite the direction of current in the next switching event.

3.5 Exploiting the stacking effect

Several options are available to exploit the stacking effect for purposes of leakage control. One obvious approach is to use a similar circuit topology to that of MTCMOS [1, 7]. Insert leakage control transistors between the power supply rails and the rest of the circuitry, but rely on the stacking effect, rather than an elevated threshold voltage to limit leakage current. Another option is to select some individual transistors and replace them by a pair of transistors with the gates tied together. Whenever such a transistor is turned off for a sufficiently long time, we will obtain a leakage reduction due to stacking effect. A third and perhaps the most attractive option is to make use of existing transistor stacks. Area penalties, performance loss, and increased switching capacitance are avoided since this does not involve adding transistors or increasing the size of pull up or pull down network;;. Except for inverters and pass gates, primitive CMOS logic gates already possess a transistor stack in either the pull down network, the pull up network, or both. R'henever a circuit is going to be idle for some length of time. it should be possible to select an input vector that maximizes the number of transistors which are turned off in each available transistor stack. If a suitable "lowleakage" input vector is not available, it may be worthwhile to alter the circuit design slightly to facilitate selection of an input vector. Recently, I-lalter and Najm [2] proposed the use of standby mode input vectors to control leakage,

but they did not identify the stacking effect as the mechanism making the leakage savings possible.

4 Conclusions

We have presented a theoretical model that predicts the quiescent leakage current and the settling time required to reach quiescent levels in transistor stack. The model is shown to correlate well with more detailed simulation results for a wide range of randomly selected design parameters. We anticipitte that the model will be useful in leakage power estimation as well as for optimizing the design of low leakage circuits.

Acknowledgements

Thanks go to Mamoon Hamid who performed countless simulations which were very useful in coming to an understanding of leakage behavior in CMOS circuits.

Appendix: derivations

Leakage of a stack of N transistors

In the steady state, the current is the same through each transistor of a stack. This assumes that other leakage currents (excluding subthreshold current) are negligible in comparison to subthreshold current. The subthreshold current through the top transistor (furthest from ground, denoted with the subscript 1) can be expressed by equation 8.

$$I_{DSq_1} = A_1 e^{\frac{1}{n\nu_T} (-(1+\gamma') \sum_{j=2}^N (V_{DSq_j}) - V_{TH_0} + \eta (V_{DD} - \sum_{j=2}^N (V_{DSq_j})))}$$
(8)
$$(1 - e^{\frac{1}{\nu_T} (V_{DD} - \sum_{j=2}^N (V_{DSq_j}))})$$

A; represents the following expression.

$$A_{i} = \mu_{0} C_{o}' x \frac{W_{eff}}{L_{eff}} (\frac{kT}{q})^{2} e^{1.8} e^{\frac{-\Delta V_{TH}}{n\nu_{T}}}$$
(9)

 V_{TH_0} is the zero bias threshold voltage. ν_T is the thermal voltage $\frac{kT}{q}$. The body effect for small values of V_S is very nearly linear. It is represented by the term $\gamma' V_S$, where γ' is the linearized body effect coefficient. η is the DIBL coefficient, representing the effect of V_{DS} ($V_{DS} = V_D - V_S$) on threshold voltage. C_{ox} is the gate oxide capacitance. μ_0 is the zero bias mobility. nis the subthreshold swing coefficient of the transistor. ΔV_{TH} accounts for variations in thresholcl voltage from one transistor to another.

The subthreshold current through the i^{th} transistor in the stack (where i > 1) is expressed by equation 10. The only difference between equations 8 and 10 is in the expression for V_{DSq_i} . For the top transistor, V_{DSq_1} can be expressed as the difference between the supply voltage and the total voltage drop across transistors lower in the stack.

$$I_{DSq_i} = A_1 e^{\frac{1}{n\nu_T} \left(-(1+\gamma') \sum_{j=i+1}^N (V_{DSq_j}) - V_{TH_0} + \eta V_{DSq_i} \right)} \left(1 - e^{\frac{1}{\nu_T} V_{DSq_i}} \right)$$
(10)

We can determine the voltage across the second transistor by equating the expressions for I_{DSq_1} and I_{DSq_2} . V_{TH_0} and all V_{DSq_1} terms for i > 2 drop out. We are left with the following expression for V_{DSq_2} . This derivation assumes that $V_{DD} >> V_{q_1}$, which proved to he true for the variety of test cases studied in this report. The derivation also takes advantage of the fact that $V_{DSq_1} >> \nu_T$ so that the $(1 - e^{\frac{1}{\nu_T}V_{DSq_1}})$ term can he ignored.

$$V_{DSq_2} = \frac{n\nu_T}{(1+2\eta+\gamma')} ln(\frac{A_1}{A_2}e^{\frac{n\nu_{DD}}{n\nu_T}} + 1)$$
(11)

The steady state voltage drop (V_{DS}) across the i^t transistor can be expipiessed in terms of the $(i-1)^{th}$ voltage drop. Equate I_{DSq_i} to $I_{DSq_{i-1}}$ and solve for V_{DSq_i} . In so doing, we obtain equation 12.

$$V_{DS_i} = \frac{n\nu_T}{(1+\gamma')} ln(1 + \frac{A_{i-1}}{A_i}(1 - e^{\frac{1}{\nu_T}V_{DS_{i-1}}}))$$
(12)

Equation 12 can be used iteratively to find the voltage drop across each transistor in the stack. V_{DSq_1} can then be obtained as $V_{DD} - \sum_{j=2}^{N} V_{DSq_j}$. Each internal node voltage can be found as the sum of voltage drops across transistors lower in the stack.

If V_{Sq_1} were to become large enough to invalidate the assumption in equation 11, then the V_{DD} term in equation 11 must be replaced by V_{DSq_1} . In this case, an iterative successive approximation approach would be required to obtain a consistent solution for V_{DSq_1} through V_{DSq_N} .

The magnitude of the steady state current can be determined using the quiescent voltage levels and the subthreshold current equation for any transistor in the stack. We choose the N^{th} transistor (bottom) for this calculation.

For the bottom transistor, V_S is equal to 0, so the current depends only upon V_{DSq_N} . This makes the calculation simpler. Furthermore, the subthreshold current is relatively insensitive to V_{DS} (in comparison to V_S).

Leakage savings ratio

If one is considering the use of a transistor stack, it may be interesting to compare the leakage current of a single transistor to the leakage current of a stack of transistors turned off. It is convenient to express this as a ratio:

$$S(N) = \frac{I_{DSq_1(1)}}{I_{DSq_1(N)}}$$
(13)

 $I_{DSq_1}(1)$ represents the quiescent leakage current of the transistor stack if only the top most transistor is turned off. In this case $V_{GS_1} = 0$. $I_{DSq_1}(N)$ represents the quiescent leakage through the top transistor if all N transistors in the stack are turned off. $V_G = 0$ for each transistor, but V_S may be greater than zero due to the stacking effect. We use equation 8 to express $I_{DSq_1(1)}$ and $I_{DSq_1}(N)$ and then plug the expressions into equation 13 to give us the savings ratio equation 14. For the transistor stack as well as a single transistor, $V_{DSq_1} \gg \nu_T$. Consequently, the $(1 - e^{\frac{1}{\nu_T}V_{DS_1}})$ is very nearly equal to one and can be dropped from the expressions for I_{DSq} . Also, since both current expressions refer to the same transistor, the A_i terms drop out (assuming that the temperature is the same in both cases).

$$S(N) = e^{\frac{1}{n\nu_T}(1+\eta+\gamma')\sum_{j=2}^N V_{DS_i}}$$
(14)

Settling time of leakage transients

Section 3.1 describes the conditions for estimating the settling time of leakage transients. In the current section, we will clarify some of the details and assumptions made in the derivation.

We estimate the time for each node to discharge as follows. During discharge, the rate by which node voltage (I/;) drops can be determined as a function of the node voltage.

$$\frac{dV_i}{dt} = -\frac{I_{dis_i}(V_i)}{C_i(V_i)} \tag{15}$$

 $I_{dis}(V_i)$ is the magnitude of the discharge current as a function of node voltage. $C_i(V_i)$ represents the node capacitance formed by the shared diffusion of the transistors above and below. C_i could include interconnect capacitance if the transistor stack is not implemented in a single contiguous strip of diffusion. C_i may also include gate and cliffusion capacitances of transistors which are not switched off. The inverse of equation 15, $\frac{dt}{dV_i}$, enables us to estimate the elapsed time corresponding to an incremental decrease of V_i . Integrating over the range by which the voltage drops, we find the time taken for the node voltage to discharge from V_{boot_i} down to the quiescent voltage level, V_{q_i} .

$$t_{dis_{i}} = \int_{V_{boot_{i}}}^{V_{q_{i}}} \frac{\delta t}{\delta V_{i}} dV_{i} = -\int_{V_{boot_{i}}}^{V_{q_{i}}} \frac{C_{i}(V_{i})}{I_{dis_{i}}(V_{i})} dV_{i} = \int_{V_{q_{i}}}^{V_{boot_{i}}} \frac{C_{i}(V_{i})}{I_{dis_{i}}(V_{i})} dV_{i}$$
(16)

Inserting expressions for $C_i(V_i)$ and $I_{dis_i}(I/;)$, the last integral for t_{dis_i} takes the form,

$$t_{dis_{i}} = \int_{V_{q_{i}}}^{V_{boot_{i}}} \frac{(C_{j0}/(1+\frac{V_{i}}{\phi_{0}})^{m}) + (C_{jsw0}/(1+\frac{V_{i}}{\phi_{0}})^{msw})}{A_{1}e^{\frac{1}{n\nu_{T}}(-(1+\gamma')V_{q_{i+1}}-V_{TH_{0}}+\eta(V_{i}-V_{q_{i+1}}))}(1-e^{\frac{-(V_{i}-V_{q_{i+1}})}{\nu_{T}}})} dV_{i} \quad (17)$$

To make this integral tractable, some simplifying assumptions are needed. We assume that the node capacitance is constant with respect to the node voltage V_i . In reality, the node capacitance (made up of diffusion or diffusion and interconnect capacitance) increases as the voltage on the node drops. To be conservative in our settling time estimate, we compute the capacitance corresponding to quiescent voltage levels. We ignore the $(1 - e^{\frac{-(V_i - V_{q_{i+1}})}{\nu_T}})$ term. The value of this term is almost exactly one until $(V_i - V_{q_{i+1}})$ approaches ν_T . The integral for t_{dis_i} now simplifies to:

$$t_{dis_{i}} = \frac{C_{i}(V_{q_{i}})}{A_{1}e^{\frac{1}{n\nu_{T}}(-(1+\gamma'+\eta)V_{q_{i+1}}-V_{TH_{0}})}} \int_{V_{q_{i}}}^{V_{boot_{i}}} e^{\frac{-\eta V_{i}}{n\nu_{T}}} dV_{i}$$
(18)

Evaluation of the integral in equation 18 leads to equation 19 for the time it takes to discharge node i.

$$t_{dis_{i}} = \frac{n C_{i} L_{eff}}{\mu_{0} C_{ox} W \nu_{T} e^{1.8} \eta} \times$$

$$e^{\frac{1}{n\nu_{T}} ((1+\gamma'+\eta)V_{q_{i+1}}+V_{TH_{0}})} \times$$

$$(e^{\frac{-\eta V_{q_{i}}}{n\nu_{T}}} - e^{\frac{-\eta V_{boot_{i}}}{n\nu_{T}}})$$

$$(19)$$

 ν_T is the thermal voltage $\frac{kT}{q}$. V_{boot} , is the voltage at internal node *i* just after switching of the transistor above, taking into account bootstrapping. V_{i} , is the quiescent level for the internal node voltage, as determined by the leakage model in section 2.1. C_i is the total capacitance of the internal node.

Since C_i decreases with voltage, we conservatively choose $C_i = C_i(V_{q_i})$. All other terms have the same definition as given in section 2.1.

The total settling time is the sum of the discharge times for each of the internal nodes of the transistor stack.

$$t_{settle} = \sum_{i=1}^{N-1} t_{dis_i} \tag{20}$$

References

- S. Shigematsu et. al. A 1-V high-speed MTCMOS circuit scheme for power-down applications. In IEEE Symposium on VLSI Circuits Digest of Technical Papers, pages 125-126, 1995.
- [2] J. P. Halter and F. Najm. A gate-level leakage power reduction method for ultra-low-power CMOS circuits. In Proceedings, IEEE Custom Integrated Circuits Conference, pages 475–478, 1997.
- [3] M.C. Jeng. Design and modeling of deep-sub-micro~neterMOSFETS. Technical Report ERL-M90/90, University of California, Berkeley, Electronics Research Laboratory, 1990.
- [4] S.-W. Kang, I<.-S. Min, and K. Lee. Parametric expression of subthreshold slope using threshold voltage parameters for MOSFET statistical modeling. IEEE *Transactions* on Electron *Devices*, 43(9):1382–1386, 1996.

- [5] Takayuki Kawahara et al. Subthreshold current reduction for decodeddriver by self-reverse biasing. *IEEE Journal of Solid-State Circuits*, 28(11):1136-1143, Nov. 1993.
- [6] A. Keshavarzi, K. Roy, and C. Hawkins. Intrinsic I_{DDQ}: Origins, reduction, and applications in deep sub-μ low-power CMOS IC's. In Proceedings IEEE International Test Conference, 1997.
- S. Mutoh et al. 1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS. *IEEE Journal of Solid-State Circuits*, 30(8):847-853, Aug. 1995.
- [8] B.J. Sheu, D.L. Scharfetter, P.K. Ko, and M.C. Jeng. BSIM: Berkeley short-channel IGFET model for MOS transistors. *IEEE Journal Solid-State Circuits*. SC-22, 1987.