Leakage Control Techniques for Designing Robust, Low Power Wide-OR Domino Logic for Sub-130nm CMOS Technologies

Bhaskar Chatterjee, Manoj Sachdev
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
bhaskar@vlsi.uwaterloo.ca

Ram Krishnamurthy*
*Microprocessor Research, Intel Labs
Intel Corporation
Hillsboro, OR, US
ram.krishnamurthy@intel.com

Abstract

In this paper, we discuss the design of leakage tolerant wide-OR domino gates for deep submicron (DSM), bulk CMOS technologies. Technology scaling is resulting in 3-5x increase in transistor $I_{OFF}/\mu m$ per generation resulting in 15%-30% noise margin degradation of high performance domino gates. We investigate several techniques that can improve the noise margin of domino logic gates and thereby ensure their reliable operation for sub-130nm technologies. Our simulations indicate that, selective usage of dual V_{TH} transistors shows acceptable energy-delay tradeoffs for the 90nm technology. However, techniques like supply voltage (V_{cc}) reduction and using non-minimum L_e transistors are required in order to ensure robust and scalable wide-OR domino designs for the 70nm generation.

1. Introduction

Aggressive technology scaling over the past 30 years has resulted in improved circuit performance and allowed designers to achieve unprecedented levels of on-die integration. However, as the transistor threshold voltage is scaled, there is a 3-5x increase in the off-state current (I_{OFF}) per generation. As a result, ensuring low power operation of complex ICs has become a major design challenge, especially for mobile and battery operated devices [1, 2, 4, 10, 14]. Figure 1 shows the scaling trends of the threshold voltage (V_{TH}) and I_{ON}/I_{OFF} ratio for both high and low V_{TH} transistors for sub 130nm technologies using the Berkeley Predictive Technology Models [3]. Our simulations indicate that, as the technology is scaled from 130nm to 70nm, the transistor I_{ON}/I_{OFF} ratio degrades by 26x for the high V_{TH} and 42x for the low V_{TH} cases. It is expected that the exponential increase in leakage current will offset the savings in switching energy *(CV^2* scaling*)* obtained from technology scaling [8, 14].
Furthermore, the degraded transistor I_{ON}/I_{OFF} ratio, scaled device geometries and power supply voltage, ever increasing switching frequency are all contributing to reduced noise margins for DSM domino logic gates. In fact, the noise margin of wide-OR domino gates is being degraded by 15%-30% per generation [11]. Such gates are normally used in the design of high performance register files (RFs) [11]. Wide-OR domino gates are especially susceptible to leakage induced false evaluations due to the presence of multiple pulldown paths. This is expected to seriously compromise their reliable operation in future DSM technologies. Thus, there exists the need to investigate techniques that can reduce leakage current and improve circuit robustness while minimizing associated performance overheads. In this paper, we investigate the following techniques in the context of wide-OR domino gates:

- Upsized p-MOS keeper [9]
- Selective usage of dual V_{TH} [5, 15]
- Pseudo-static technique [11]
- Selective usage of non-minimum L_e transistors [7, 16]
- Supply voltage reduction [6, 13]

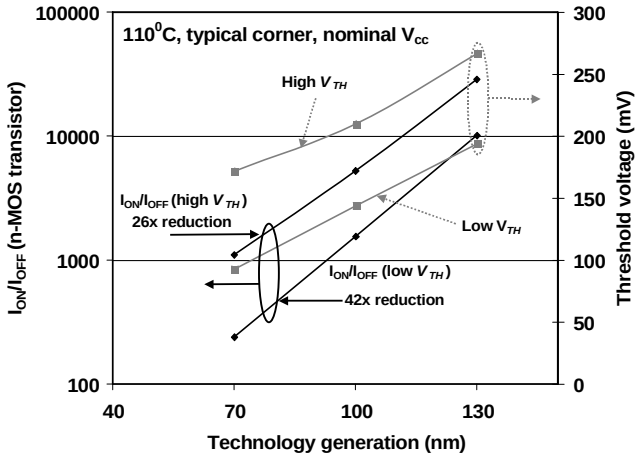

Figure 1: I_{ON}/I_{OFF} and V_{TH} scaling for sub 130nm generations

We study the impact of the above techniques on the following parameters: propagation delay, leakage and switching energy, and DC robustness. The rest of the paper is organized as follows: in Section 2 we discuss the design of wide-OR domino gates and quantify the DC robustness degradation caused by technology scaling. In Sections 3 and 4, we present the different techniques, and their associated design tradeoffs for the 90nm and 70nm technologies. Section 5 is for conclusions.

2. Wide Domino: Design and Robustness Scaling

Wide-OR domino gates are used in the design of local and global bit lines (LBL, GBL) of high performance RFs. Figure

2 shows an 8-wide domino gate with 2-stack n-MOS pulldown implemented using the compound domino logic (CDL). In addition to the 2-stack pulldowns, high performance functional unit blocks (FUBs) also use single n-MOS pulldowns (GBLs). The inputs to the pulldown network are normally domino compatible. This allows removal of the clocked footer transistor, reduces the stack height, improves performance and lowers switching energy.

In this paper, we consider the worst-case conditions for both DC robustness and propagation delay. As indicated in Figure 2, the worst-case gate delay occurs when only one of the pulldown paths is selected and the wide-OR gate operates as a high performance MUX. During the evaluation phase (CLK=1), if the gate signals of both transistors are high (A_0, B_0=1), the dynamic node evaluates to ground (Dyn_node=0) resulting in the static gate output transitioning to V_{cc} (OUT=1). Typically, in RF applications, the signals B_0-B_7 are setup ahead of time while the MUX select signals (A_0-A_7) are timing critical [11]. This fact will subsequently be exploited in the selective assignment of dual V_{TH} and non-minimum L_e for the 90nm and 70nm designs.

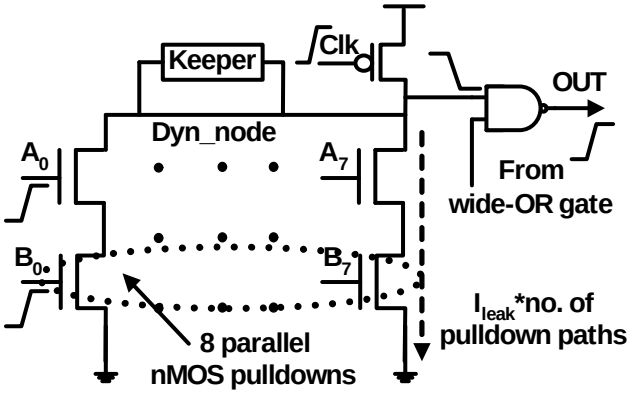

Figure 2: Wide-OR domino gate for RFs (LBL organization)

In this paper, we consider DC robustness as our metric for determining noise margin of wide-OR domino gates. The DC robustness is defined with respect to the node OUT (for both 2-nMOS→LBL, and 1-nMOS→GBL pulldowns) and can be better understood with the help of the simulation waveforms shown in Figure 3. DC robustness waveforms are obtained under worst-case leakage conditions when the signals A_0-A_7 are subjected to DC noise (simulated using a slow ramp signal). The voltage when the wide-OR domino output (OUT) equals the input, is identified as the unity gain noise margin (UGNM) point. DC robustness for a given technology is defined as the normalized UGNM ($UGNM/V_{cc}$). This definition for DC robustness (UGNM) is well established in the context of leakage tolerant domino logic design [9, 11, 12].

The results shown in Figure 3 indicate that, a 5% p-MOS keeper results in DC robustness of ~17% for an 8-wide domino gate for the 130nm technology under worst-case conditions. We use this as our reference design to set the target DC robustness for the 90nm and 70nm technologies. This allows us to compare the different techniques and quantify various design tradeoffs. It is possible to set a different absolute value for the robustness threshold, but the general trends and energy-delay tradeoffs would still remain unaffected.

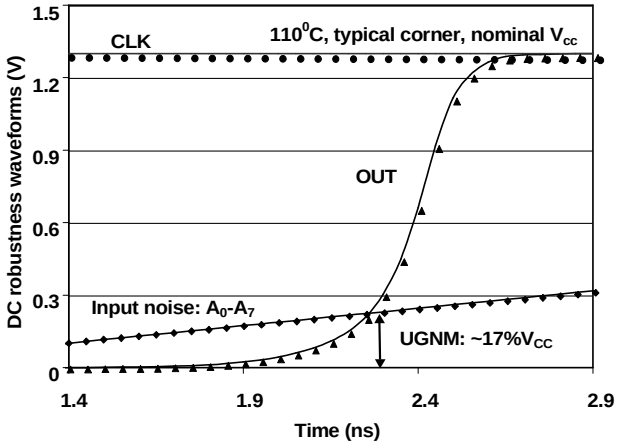

Figure 3: DC robustness waveforms for 130nm

Figure 4 shows the impact of technology scaling on DC robustness for the 8-wide, LBL with 5% p-MOS keeper. Our results indicate that, for the 90nm (70nm) technology, there is 24% (41%) degradation in DC robustness. It should be noted that the data in Figure 4 for the 130nm and 90nm technologies, correspond to all low-V_{TH} designs. On the other hand, the data for 70nm corresponds to a dual V_{TH} design. This is because an all-low V_{TH} 70nm design shows unacceptable noise margin under worst-case conditions and fails to operate due to excessive transistor leakage. The DC robustness for wide-OR domino gates with 1-nMOS pulldown also shows similar scaling trends as those in Figure 4. It is clear from these results that, the 3-5x increase in I_{OFF} current per generation will significantly degrade the noise margin of high performance domino logic gates resulting in possible false evaluations. Therefore, we need to explore alternate design/leakage control techniques that improve DC robustness and allow reliable operation of DSM domino gates.

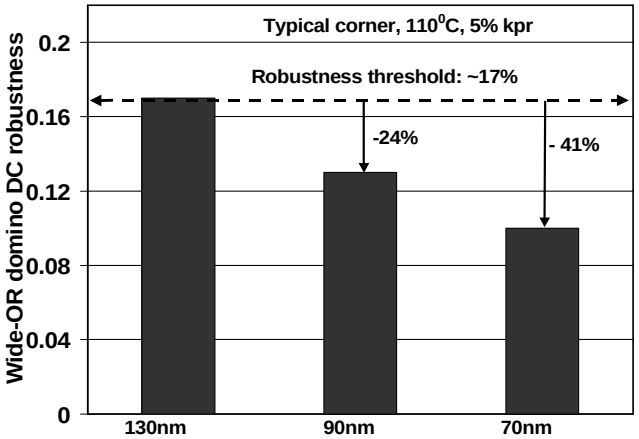

Figure 4: Wide-OR domino DC robustness scaling trends

3. Techniques for Improving Robustness

In this section we discuss some of the different techniques that can be used to improve the UGNM and robustness of wide-OR domino gates for DSM technologies. We present the energy-delay tradeoffs associated with the techniques mentioned earlier, discuss their applicability to both 2-stack and 1-stack domino designs (LBL and GBL) and show their scaling trends for the 90nm and 70nm generations.

3.1 Keeper Upsizing

The simplest technique to improve domino logic noise margin is to strengthen the p-MOS pullup keeper. This ensures that the normally ON p-MOS transistor sources a larger linear mode current to offset the increased I_{OFF} current of the pulldown network. Our simulations indicate that, the p-MOS keeper has to be upsized by 2x (2.3x) for the 90nm (70nm) generations to maintain iso-robustness (UGNM ~17%). As the keeper size is increased, it contends with the pulldown network, resulting in increased propagation delay and switching energy. Figure 5 shows the energy-delay tradeoffs for an 8-wide 2-stack LBL design for the 90nm and 70nm generations using upsized keepers. Our results indicate that, when upsized keepers are used to meet the noise margin threshold, there is a 12%-16% delay degradation, and ~2% increase in switching energy. In addition, there is an 11%-14% reduction in leakage energy. This results from the fact that the dynamic node is firmly anchored to V_{cc} (reduced DC droop) causing less subthreshold leakage in the subsequent static NAND gate. This technique is simple and can be used for domino gates with both 2-stack and 1-stack (LBL, GBL) n-MOS pulldowns. However, it is clear that the energy-delay tradeoffs associated with keeper upsizing are not favourable for designing high performance datapaths.

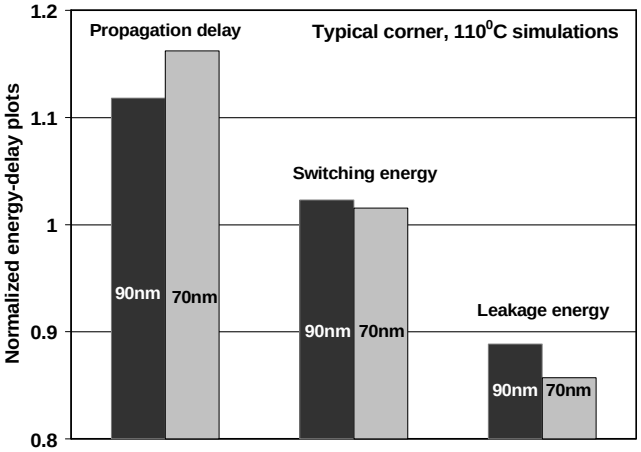

Figure 5: Impact of upsized keeper on DSM domino gates

3.2 Dual V_{TH} Technique

The dual-V_{TH} technique is based on the selective usage of low and high threshold transistors to minimize leakage current while limiting the delay degradation. The high V_{TH} transistors help in the reduction of leakage current and charge loss from the dynamic node thereby improving the UGNM. The 2-stack LBL domino gates are organized such that the gate signal for the bottom transistors B_0-B_7 are connected to the local bitcells and are setup ahead of time. However, the performance critical Read Select signals typically drive long interconnects and are connected to the transistors A_0-A_7. Under worst-case conditions, these signals may be subjected to input noise while signals B_0-B_7, are held at V_{cc} and are ON. Consequently, transistors A_0-A_7 determine the domino gate leakage and worst-case UGNM. In the dual-V_{TH} scheme, we use high V_{TH} for these transistors, while low V_{TH} transistors are used for B_0-B_7 to limit the overall performance degradation. Figure 6 shows the simulation results indicating the energy-delay tradeoffs involved with a dual-V_{TH} LBL scheme for the 90nm technology.

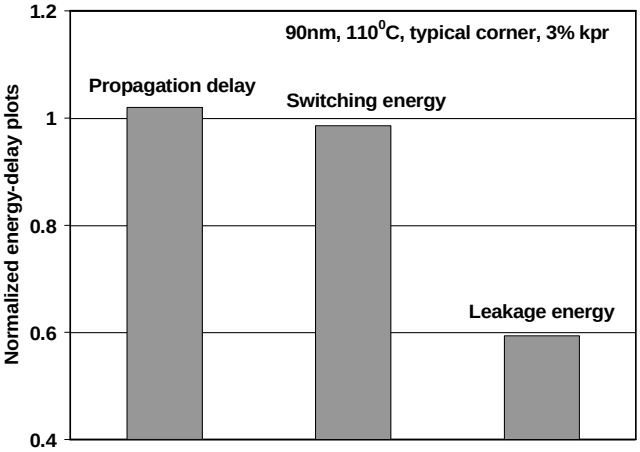

Figure 6: Dual V_{TH} domino logic energy-delay tradeoffs for 90nm

Our results indicate that, the reduction in leakage current associated with the dual-V_{TH} technique, allows us to use a weaker p-MOS keeper (3%) to meet the noise margin threshold. Therefore, for the 90nm technology, it is possible to limit the delay degradation to within 2%. The selective usage of high V_{TH} transistors also allows 41% reduction in leakage energy. In addition, the weaker p-MOS keeper results in less pulldown contention allowing a 1.5% savings in switching energy.

However, for the 70nm technology, the leakage current of both the high and low V_{TH} transistors increase by 3-5x. As a result, the dual-V_{TH} technique needs to be used in conjunction with upsized p-MOS keeper to meet the robustness threshold. Therefore, to maintain iso-robustness, a dual-V_{TH} LBL design needs 2.3x (11.3%) p-MOS keeper, which results in 16% delay degradation. Further more, the dual-V_{TH} technique cannot be used effectively for designing robust 1-stack wide domino gates. Thus, GBL designs require an all-high V_{TH} pulldown with a 1.9x (9.5%) upsized keeper resulting in 10% delay degradation. In both cases, the upsized keeper results in ~2% increase in switching energy due to extra contention during evaluation. Thus, it is clear from the above results that, for the 70nm generation, the dual-V_{TH} technique alone, cannot guarantee robust operation of wide-OR domino logic gates.

0-7695-2093-6/04 $20.00 © 2004 IEEE

3.3 Pseudo-Static Technique

The pseudo-static technique [11] has been advanced as a means for designing robust wide-OR domino logic gates for DSM technologies. In this section we briefly study this technique and discuss its impact on LBL, GBL designs. The pseudo-static circuit technique is explained with the help of Figure 7. This technique improves the UGNM by reducing the leakage current and dynamic node charge loss through transistors N2-N16. Firstly, the order of the pulldown n-MOS transistors is reversed, whereby the performance critical signals (A_0-A_7) are connected to the bottom of the LBL stack. Secondly, a minimum sized p-MOS transistor (P1) is used to pullup the internal stack node voltage (V_X) to V_{cc} for all deselected paths.

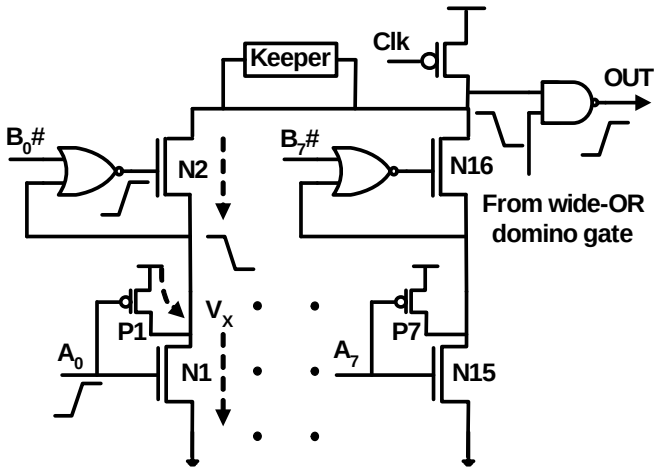

Figure 7: Robust domino design using pseudo-static scheme

Thirdly, a 2 input static NOR gate is used to turn OFF transistor N2 in case the pulldown path is deselected (A_0=0). This scheme ensures that both transistors in the n-MOS stack are OFF, N2 has a higher "effective" threshold voltage (reverse body bias and reduced DIBL effect) and a negative V_{GS} bias voltage. As a result, there is significant reduction in leakage current though N2, resulting in improved UGNM. In fact, our simulations indicate that it is possible to maintain iso-robustness for the 70nm technology, while using an all low V_{TH} n-MOS pulldown and 3% p-MOS keeper.

However, the above technique suffers from several drawbacks that result in delay degradation, and increased overall switching and leakage energy:

1. The reversal of transistor order results in performance critical signals (A_0-A_7, Read Selects) being placed further from the gate output.
2. The p-MOS transistor (P1-P7) adds additional capacitance to the intermediate node V_X and precharges the node to V_{cc}. This is unlike the normal LBL design where the data is setup ahead of time, pre-discharging the corresponding node to ground.
3. The critical path has an extra stage of inversion due to the 2-input NOR gate. Further more, the NOR gate has to be designed in order to aid the 0→1 transition, resulting in increased p-MOS transistor widths. As a result, there is increased leakage through the deselected NOR gates and added capacitive loading at the intermediate node V_X.
4. When a particular pulldown path is deselected (A_0=0), the pMOS transistor (P1) turns ON, and the voltage across N1 (V_X) approaches V_{cc}. The final steady-state voltage is reached when the I_{OFF} current of N2 and linear current of P1 equal the I_{OFF} of N1. Our simulations for the 70nm technology indicate that, under worst-case conditions, the V_X node voltage equals ~$0.95V_{cc}$. This implies that even though the leakage current through N2 is reduced resulting in improved UGNM, the overall leakage current is actually increased, with the extra current flowing through the parallel path formed by transistors P1-N1.
5. The extra capacitance introduced by P1-P7 and NOR gates result in higher switching energy.
6. This technique depends on the availability of the intermediate node V_X and is therefore not suitable for robust GBL designs with single n-MOS pulldown stacks.

The above drawbacks associated with the pseudo-static technique, offset the delay improvements resulting from an all low V_{TH} pulldown and 3% p-MOS keeper design. This is clear from the energy-delay tradeoffs for the 70nm LBL design as shown in Figure 8. Our simulations indicate that, the pseudo-static LBL meets the DC robustness threshold, while resulting in a 9% delay penalty. In addition, there is an 8% increase in switching energy, with 4% savings in leakage energy. This implies that the static-NOR delay and leakage (2 p-MOS stack upsized for improved performance) degrade the overall switching and leakage energy of the wide-OR domino gate. In addition, the worst-case noise margin for 1-stack n-MOS pulldown degrades with scaling and cannot be improved using this circuit technique.

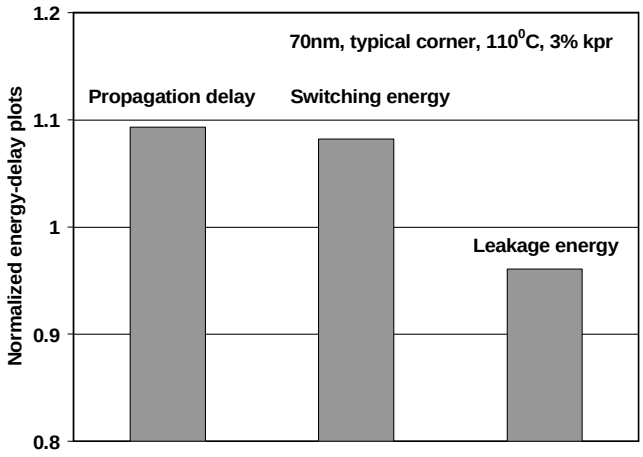

Figure 8: Pseudo-static LBL energy-delay plots for 70nm technology

4. Non-minimum L_e, Scaled V_{cc}: Robust 70nm design

In this section, we focus on the selective usage of non-minimum channel length (L_e) transistors and supply voltage

scaling on wide-OR domino gates for the 70nm generation. We first investigate the effect of both these techniques on the I_{ON}-I_{OFF} plane at the transistor level, and then discuss the energy-delay tradeoffs associated with both LBL (2-stack) and GBL (1-stack) organizations.

4.1. Transistor Level I_{ON}-I_{OFF} Tradeoffs

There are several different techniques that can be used to reduce transistor leakage current. Among these techniques, some depend on supply voltage reduction, while others are based on increasing the transistor threshold. The reduction of power supply has a twofold impact on leakage power: there is a reduction in transistor DIBL current and lowering of the $V_{cc}.I_{OFF}$ product. On the other hand, increasing the transistor channel length results in higher threshold voltage. This in turn results in an exponential reduction of the weak inversion current. However, both of these techniques also result in reduced transistor I_{ON} current [$\propto (V_{cc} - V_{TH})^{\alpha}$] and cause performance degradation. A technique that offers larger leakage power/energy reductions with minimum delay degradation is more efficient and is suitable for robust, high performance logic designs. Figure 9 compares the effectiveness of two techniques for the 70nm technology using transistor level simulations when the supply voltage is reduced by 25%, and the channel length is increased by 33%, respectively. We compare these two techniques in the [$V_{cc}.I_{OFF}$]-[V_{cc}/I_{ON}] plane. The first term is the leakage power while the second term reflects the delay degradation associated with each technique.

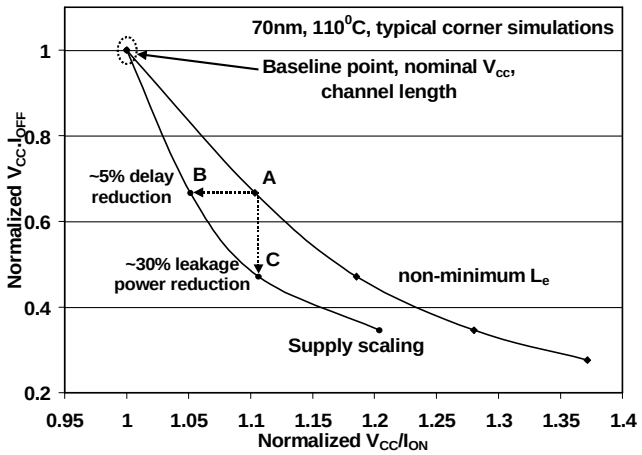

Figure 9: Leakage techniques compared for 70nm technology

Our simulation results indicate that, lowering the power supply is a more efficient leakage control technique than using non-minimum L_e since it results in less delay degradation. It is clear from data points A and B, that for the same amount of leakage power, supply scaling offers 5% less delay degradation. Conversely, for the same delay (points A and C), there is ~30% lower leakage power consumption. In addition, there is a quadratic savings in switching energy resulting from supply voltage scaling as opposed to a near linear increase associated with using non-minimum channel length transistors. This increase can be attributed to an increase in switching capacitance due to higher effective $W.L_e$ product of the transistors.

4.2. Robust, Energy Efficient 70nm Wide-OR Domino

In this section, we study the impact of the above techniques on 8-wide, 2-stack pulldown 70nm domino logic gates. Both these techniques are also applicable to 1-stack n-MOS pulldown (GBL) domino designs. In this study, the domino supply voltage was lowered up to 28%. The channel lengths of transistors (A_0-A_7) were increased (up to 33%) while those at the bottom (B_0-B_7) were left unchanged. This is similar to the approach adopted for the dual-V_{TH} design as described earlier in Section 3.2.

Figure 10 shows the impact of non-minimum L_e transistors on LBL designs while meeting the noise margin threshold at each data point. As the channel length is increased, the leakage current reduces allowing downsizing of the p-MOS keeper (11.3%→6%). It is clear from these results, that the reduction in leakage energy is compensated for by an increase in switching energy. Therefore, the reduction in total energy depends on the relative ratio of the switching and leakage energy components. In addition, the reduction in I_{OFF} depends on the proportion of the weak inversion current in the total off-state current. Our results indicate that, with the selective usage of non-minimum L_e transistors (L_e+33%), the propagation delay degrades by ~4% while resulting in ~2% savings in total energy. It should be noted that the weakened keeper helps limit the delay impact associated with this technique to within 4%.

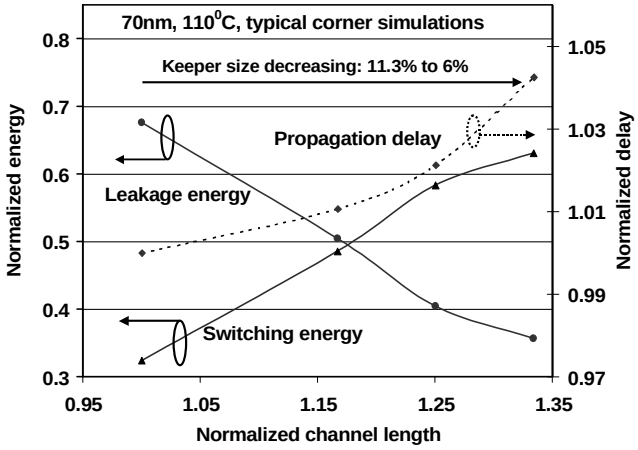

Figure 10: Energy-delay plots for 70nm using non-min. L_e

The results in Figure 11 correspond to the case when the supply voltage is reduced from the nominal value to $0.72V_{cc}$. All the data points correspond to 17% DC noise margin. As the supply voltage is scaled, there is a

corresponding reduction in leakage current allowing the p-MOS keeper to be downsized from 11.3% to 5%.

Our results indicate that when the power supply is scaled by 14%, the delay degradation is ~4% allowing ~35% reduction in total energy. This implies that limited supply voltage scaling can be used for DSM wide-OR domino logic gates to ensure robust designs and low power operation while limiting performance penalty to within acceptable limits. A similar 14% scaling of the power supply for the GBL results in ~5% delay degradation with 38% savings in total energy.

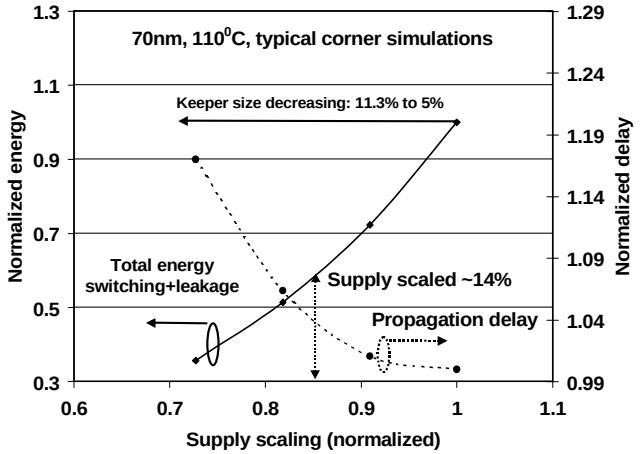

Figure 11: Energy-delay plots for 70nm with supply scaling

5. Conclusion

In this paper, we discussed the impact of technology scaling on domino logic gates. In particular, we focussed on the noise margin degradation of wide-OR domino gates. We compared several different circuit and leakage control techniques that can be used to ensure robust domino logic operation for the sub-130nm generations. Our results indicate that while dual-V_{TH} technique is suitable for the 90nm technology, limited supply voltage scaling (10%-15%) followed by usage of non-minimum L_e transistors demonstrate improved energy-delay tradeoffs for the 70nm generation. It is expected that such techniques will ensure robust, low-power operation of high performance DSM domino logic gates.

6. Acknowledgements

Authors would like to acknowledge O. Semenov, S. Naraghi and C. Kwong from the University of Waterloo, and S. Hsu and S. Borkar from Intel Corp. for encouragement and support.

7. References

[1] J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proceedings of the IEEE*, vol. 83, no. 4, pp. 619-635, 1995.

[2] A. P. Chandrakasen, S. Sheng, and R. W. Brodersen, "Low power CMOS Digital Design," *IEEE Journal of Solid State Circuits*, vol. 27, no. 4, pp. 473-484, 1992.

[3] http://www-device.eecs.berkeley.edu: BSIM3 100nm and 70nm predictive technology process files.

[4] V. De, and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," *Proceedings of the International Symposium on Low Power Design*, pp. 163-168, 1999.

[5] K. Roy, S. Mukhopadhyay, and H. M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305-327, Feb. 2003.

[6] M. R. Stan, "Optimal Voltages and Sizing for Low Power," *12th IEEE International Conference on VLSI Design*, pp. 428-433, 1999.

[7] N. Sirisantana, L. Wei, and K. Roy, "High-Performance Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide Thickness," *Proceedings of the International Conference on Computer Design*, pp. 227-232, 2000.

[8] T. Kuroda, "CMOS Design Challenges to Power Wall," *International Conference on Microprocessors and Nanotechnology*, pp. 6-7, 2001.

[9] S. O. Jung, K. W. Kim, and S. Kang, "Noise Constrained Power Optimization for Dual V_T Domino Logic," *Proceedings of the International Symposium on Circuits and Systems*, pp.158-161, 2001.

[10] A. Chandrakasan, W.J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*. IEEE Press, Piscataway, N.J., 2000.

[11] R. Krishnamurthy, A. Alvandpour, G. Balamurugan, N. Shanbag, K. Soumyanath, and S. Borkar, "A 130nm 6-GHz 256x32 bit Leakage-Tolerant Register File," *IEEE Journal of Solid State Circuits*, vol. 37, no. 5, pp. 624-632, May 2002.

[12] S. Thompson, I. Young, and M. Bohr, "Dual Threshold and Substrate Bias: Keys to High Performance, Low Power, 0.1μm Logic Designs," *Symposium on VLSI Technology*, pp. 69-70.

[13] R. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. Chatterjee, M. Sachdev, and S. Borkar, "Dual supply voltage clocking for 5GHz 130nm integer execution core", *Symposium on VLSI Circuits*, pp. 128-129, 2002.

[14] T. Kuroda, "Low-Power, High Speed CMOS VLSI Design," *Proceedings of the IEEE Conference on Computer Design*, pp. 310-315, 2002.

[15] J. T. Kao, and A. Chandrakasen, "Dual-Threshold Voltage Techniques for Low-Power Digital Circuits," IEEE Journal of Solid State Circuits, vol. 35, no. 7, pp. 1009-1018, July 2000.

[16] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy and S. Borkar, "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies," *Proceedings of the International Symposium of Low Power Electronics and Design*, pp. 122-127, 2003.