Scaling of Stack Effect and its Application for Leakage Reduction

Siva Narendra[¶§], Shekhar Borkar[§], Vivek De[§], Dimitri Antoniadis[¶], and Anantha Chandrakasan[¶]

[¶]Microsystems Technology Laboratories
Massachusetts Institute of Technology, Cambridge, MA, 02139
{naren, anantha, daa}@mtl.mit.edu

[§]Microprocessor Research Laboratories
Intel Corporation, Hillsboro, OR, 97124
{shekhar.y.borkar, vivek.de}@ intel.com

ABSTRACT

Technology scaling demands a decrease in both V_{dd} and V_t to sustain historical delay reduction, while restraining active power dissipation. Scaling of V_t however leads to substantial increase in the sub-threshold leakage power and is expected to become a considerable constituent of the total dissipated power. It has been observed that the stacking of two *off* devices has smaller leakage current than one *off* device. In this paper we present a model that predicts the scaling nature of this leakage reduction effect. Device measurements are presented to prove the model's accuracy. Use of stack effect for leakage reduction and other implications of this effect are discussed.

1. INTRODUCTION

To limit the energy and power increase in future CMOS technology generations, the supply voltage (V_{dd}) will have to continually scale [1]. The amount of energy reduction depends on the magnitude of V_{dd} scaling [2]. Along with V_{dd} scaling, the threshold voltage (V_t) of MOS devices will have to scale to sustain the traditional 30% gate delay reduction. These V_{dd} and V_t scaling requirements pose several technology and circuit design challenges [3-5].

One such challenge is the rapid increase in sub-threshold leakage power due to V_t scaling. Should the present scaling trend continue it is expected that the sub-threshold leakage power will become a considerable constituent of the total dissipated power [6]. In such a system it becomes crucial to identify techniques to reduce this leakage power component. It has been shown previously that the stacking of two *off* devices has significantly reduced sub-threshold leakage compared to a single *off* device [7-9]. This concept of stack effect is illustrated in Figure 1. In rest of the paper the term leakage refers to sub-threshold leakage.

In this paper we present a model that predicts the stack effect factor, which is defined as the ratio of the leakage current in one *off* device to the leakage current in a stack of two *off* devices. Model derivation based on device fundamentals and verification of the model through statistical device measurements from 0.18μm and 0.13μm technology generations are presented in Section 2. The scaling nature of the stack effect leakage reduction factor is also discussed in the next section.

ISLPED'01, August 6-7, 2001, Huntington Beach, California, USA.

One solution to the problem of ever-increasing leakage is to force a non-stack device to a stack of two devices without affecting the input load, as shown in Figure 2. By ensuring iso-input load, the previous gate's delay and the switching power will remain unchanged. Logic gates after stack forcing will reduce leakage power, but incur a delay penalty, similar to replacing a low- V_t device with a high-V_t device in a dual-V_t design [10]. In a dual-V_t design the low-V_t devices are used in performance critical paths and the high-V_t devices in the rest [11]. Usually a significant fraction of the devices can be high-V_t or forced-stack since a large number of the paths are non-critical. This will reduce the overall leakage power of the chip without impacting operating clock frequency. In Section 3 we discuss the stack forcing method to reduce leakage in paths that are not performance critical. This stack forcing technique can be either used in conjunction with dual-V_t or can be used to reduce the leakage in a single-V_t design. Differences between achieving leakage reduction through forced-stacks and channel length increase are discussed in Section 4. Conclusions and future work are described in Section 5.

2. MODEL FOR STACK EFFECT FACTOR

Let I_1 be the leakage of a single device of unit width in *off* state with its $V_{gs} = V_{bs} = 0$ V and $V_{ds} = V_{dd}$. If the gate-drive, body bias, and drain-to-source voltages reduce by ΔV_g, ΔV_b, and ΔV_d respectively from the above mentioned conditions, the leakage will reduce to,

$$I'_1 = I_1 \quad 10^{-\frac{1}{S}\left[\Delta V_g + \lambda_d \Delta V_d + k_\gamma \Delta V_b\right]}$$

where S is the sub-threshold swing, λ_d is the drain-induced barrier lowering (DIBL) factor, and k_γ is the body effect coefficient. The above equation assumes that the resulting $V_{ds} > 3kT/q$ [6]. For a two-device stack shown in Figure 3, a steady state condition will be reached when the intermediate node voltage V_{int} approaches V_x such that the leakage currents in the upper and lower devices are equal.

Under this condition, the leakage currents in the upper and lower devices can be expressed as,

$$I_{\text{stack-}u} = w_u I_1 \quad 10^{\frac{-(1+\lambda_d+k_\gamma)V_x}{S}}$$

$$I_{\text{stack-}l} = w_l I_1 \quad 10^{\frac{-\lambda_d(V_{dd}-V_x)}{S}}$$

and the intermediate node voltage will be,

$$V_x = \frac{\lambda_d V_{dd} + S\log\frac{w_u}{w_l}}{1+k_\gamma+2\lambda_d}$$

For short channel devices the body terminal's control on the channel is negligible compared to gate and drain terminals, implying $k_\gamma << 1 + 2\lambda_d$. Hence the steady state value, V_x, of the intermediate node voltage can be approximated as,

$$V_x \approx \frac{\lambda_d V_{dd} + S \log \frac{w_u}{w_l}}{1+2\lambda_d}$$

Substituting V_x in either $I_{stack\text{-}u}$ or $I_{stack\text{-}l}$ will yield the leakage current in a two-stack given by,

$$I_{stack} = w_u^{\alpha} w_l^{1-\alpha} I_1 \; 10^{\frac{-\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$\text{where} \quad \alpha \approx \frac{\lambda_d}{1+2\lambda_d}$$

The leakage reduction achievable in a two-stack comprising of devices with widths w_u and w_l compared to a single device of width w is given by,

$$X = \frac{I_{device}}{I_{stack}} = \frac{w}{w_u^{\alpha} w_l^{1-\alpha}} 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$= 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)} \quad \text{when} \quad w_u = w_l = w$$

The stack effect factor, when $w_u = w_l = w$, can be rewritten as,

$$X = 10^{\frac{\lambda_d V_{dd}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)} = 10^U$$

where U is the universal two-stack exponent which depends only on the process parameters, λ_d and S, and the design parameter, V_{dd}. Once these parameters are known, the reduction in leakage due to a two-stack can be determined from the above model. It is essential to point out that the model assumes the intermediate node voltage to be greater than $3kT/q$.

To confirm the model's accuracy we performed device measurements on test structures fabricated in 0.18µm and 0.13µm process technologies. Results discussed in the rest of the section are from NMOS device measurements, but similar results hold true for PMOS devices as well.

Figure 4 shows NMOS device measurements under different temperature, V_{dd}, body bias, and channel length conditions for 0.18µm technology generation, which prove the accuracy of the theoretical model. It is important to note that the model discussed above doesn't include the impact of diode junction leakages that originate at the intermediate stack node. In Figure 4 the model's accuracy deviates the most under reverse body bias for nominal channel length devices, where the ratio of diode junction leakage to sub-threshold leakage current increases.

It is known that the stack effect factor strongly depends on λ_d as suggested by the model. Also a decrease in the channel length (L) will increase λ_d in a given technology [12]. So, any increase in the leakage of a single device due to decrease in L will not increase leakage of a two-stack at the same rate. This is illustrated in Figure 5 where increase in two-stack leakage is at a slower rate than that of a single device.

Figure 6 illustrates the average stack effect factor for the nominal channel devices in both 0.18µm and 0.13µm technology generations obtained from both the measurements and the model. The increase in stack effect factor at a given V_{dd} with technology scaling is attributed to increase in λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13µm technology generation is due to the same effect.

In 0.13µm generation, the low-V_t device will dominate chip leakage. Figure 7 shows the scaling of stack effect from a 0.18µm device to a 0.13µm low-V_t device based on device measurements under different V_{dd} scaling scenarios. Since λ_d is expected to increase due to worsening device aspect ratio and since V_{dd} scaling will slow down due to related challenges [13], stack effect leakage reduction factor is expected to increase with technology scaling. The predicted scaling of stack effect factor from 0.18µm to 0.06µm is depicted in Figure 8.

This scaling nature of stack effect factor makes it a powerful technique for leakage reduction in future technologies. In the next section we describe a circuit technique for taking advantage of stack effect to reduce leakage at a functional block level.

3. LEAKAGE REDUCTION USING FORCED-STACKS

As shown earlier, stacking of two devices that are *off* has significantly reduced leakage compared to a single *off* device. However due to the iso-input load requirement and due to stacking of devices, the drive current of a forced-stack gate will be lower resulting in increased delay. So, stack forcing can be used only for paths that are non-critical, just like using high-V_t devices in a dual-V_t design [10-11]. Forced-stack gates will have slower output edge rate similar to gates with high-V_t devices. Figure 9 illustrates the use of techniques that provide delay-leakage trade-off. As demonstrated in the figure, paths that are faster than required can be slowed down which will result in leakage savings. Such trade-offs are valid only if the resulting path still meets the target delay. Figure 10 shows the delay-leakage trade-off due to n-stack forcing of an inverter with fan-out of 1 under iso-input load conditions in a dual-V_t 0.13µm technology [14].

By properly employing forced-stack one can reduce standby and active leakage of non-critical paths even if a dual-V_t process is not available. This method can also be used in conjunction with dual-V_t. Stack forcing provides wider coverage in the delay-leakage trade-off space as illustrated in Figure 10.

Functional blocks have naturally stacked gates such as NAND, NOR, or other complex gates. By maximizing the number of natural stacks in *off* state during standby by setting proper input vectors, the standby leakage of functional block can be reduced. Since it is not possible to force all natural stacks in the functional block to be in *off* state the overall leakage reduction at a block level will be far less than the stack effect leakage reduction possible at a single logic gate level [7]. With stack forcing the potential for leakage reduction will be higher. Figures 11(a) and 11(b) illustrates such an example.

Forcing a stack in both n- and p-networks of a gate will guarantee leakage reduction due to stacking, independent of the input logic level. Such an example is shown in Figure 11(c). To reiterate, stack forcing can be applied to paths only if increase in delay due to stacking does not violate timing requirements. Gates that can force stack effect independent of its input vectors

will automatically go into leakage reduction mode when the intermediate node of the stack reaches the steady state voltage. This will boost standby and active leakage reduction since no specific input vector needs to be applied.

4. STACK EFFECT VS. CHANNEL LENGTH INCREASE

It is possible to facilitate delay-leakage trade-off by increasing the channel length of devices [15] that are in non-critical paths. To maintain iso-input load the channel width will have to be reduced along with increase in the channel length. Figure 12, shows the mean leakage reduction achievable by increasing the channel length. Mean leakage is defined as the geometric mean of the leakages with and without variation in critical dimension around the channel length of interest. This mean leakage is expected to model the leakage of a chip that has within-die variation in critical dimension. In Figure 12 the channel length of interest is given by η x 0.18 μm and stack leakage is for a stack of two devices with η of 1 and $w_u=w_l=½w$. As it is clear from Figure 12, the channel length has to be increased 3 times as that of the nominal channel length to match the mean leakage of a two-stack of 0.18μm devices. The main reason for such a large increase is attributed to the reverse short channel effect that is present due to halo doping [13] where V_t reduces with increase in channel length.

Figure 13 shows the energy-delay trade-off of an inverter under different configurations with fan-out of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than delay due to increasing channel length, especially when compared to iso-standby leakage ($\eta \approx 3$) configuration. As summarized in Figure 14, η of 2 has about the same delay as that of the two-stack with η of 1 but with a 2.3X higher mean leakage. On the other hand η of 3 provides about the same mean leakage as the two-stack but with 60% higher delay.

5. CONCLUSIONS

We presented a model based on device fundamentals that predicted the scaling nature of stack effect based leakage reduction. Device measurements verified the model's accuracy across different temperature, channel length, body bias, supply voltage, and process technology. Modes for using stack forcing to reduce standby and active leakage components were discussed and the advantage of stack forcing over channel length increase for delay-leakage trade-off was demonstrated. Stack forcing assignment for standby and active leakage reduction at a functional block level with and without dual-V_t will be explored in the future.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge A. Keshavarzi and J. Tschanz for discussions and B. Bloechel for assistance in measurements.

7. REFERENCES

[1] S. Borkar, "Technology Trends and Design Challenges for Microprocessor Design", *ESSIRC*, pp. 7-8, Sep. 1998.

[2] A. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital design", *IEEE J. Solid-State Circuits*, vol. 27, pp. 473-484, Apr. 1992.

[3] D. Antoniadis and J. E. Chung, "Physics and Technology of Ultra Short Channel MOSFET Devices", *Intl. Electron devices Meeting*, pp. 21-24, 1991.

[4] V. De and S. Borkar, "Technology and Design Challenges for Low Power & High Performance", *Intl. Symp. Low Power Electronics and Design*, pp. 163-168, 1999.

[5] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, "CMOS Technology Scaling for Low Voltage Low Power Applications", *IEEE Symp. Low Power Elec.*, pp. 56-57, 1994.

[6] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*, IEEE Press, pp. 46-47, 2000.

[7] Y. Ye, S. Borkar, and V. De, "A Technique for Standby Leakage Reduction in High-Performance Circuits," *Symp. of VLSI Circuits*, pp. 40-41, 1998.

[8] J. P. Halter and F. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits," *Custom Integrated Circuits Conf.*, pp. 475-478, 1997.

[9] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," *Intl. Symp. Low Power Electronics and Design*, pp. 239-244, 1998.

[10] L. Su, R. Schulz, J. Adkisson, K. Beyer, G. Biery, W. Cote, E. Crabbe, D. Edelstein, J. Ellis-Monaghan, E. Eld, D. Foster, R. Gehres, R. Goldblatt, N. Greco, C. Guenther, J. Heidenreich, J. Herman, D. Kiesling, L. Lin, S-H. Lo, McKenn, "A high-performance sub-0.25μm CMOS technology with multiple thresholds and copper interconnects," *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 18-19, 1998.

[11] D. T. Blaauw, A. Dharchoudhury, R. Panda, S. Sirichotiyakul, C. Oh, and T. Edwards "Emerging power management tools for processor design," *Intl. Symp. Low Power Electronics and Design*, pp. 143-148, 1998.

[12] Z. Liu, C. Hu, J. Huang, T. Chan, M. Jeng, P. Ko, and Y. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Trans. Elec. Devices*, vol. 40, no. 1, pp. 86-95, January 1993.

[13] Y. Taur, "CMOS Scaling beyond 0.1μm: how far can it go?" *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 6-9, 1999.

[14] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Bradenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler, S. Thompson, B. Tufts, J. Xu, S. Yang, and M. Bohr, "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects," *Intl. Elec. Devices Meeting*, pp. 567-570, December 2000.

[15] D. Dobberpuhl, "The Design of a High Performance Low Power Microprocessor," *Intl. Symp. Low Power Electronics and Design*, pp. 11-16, 1996.

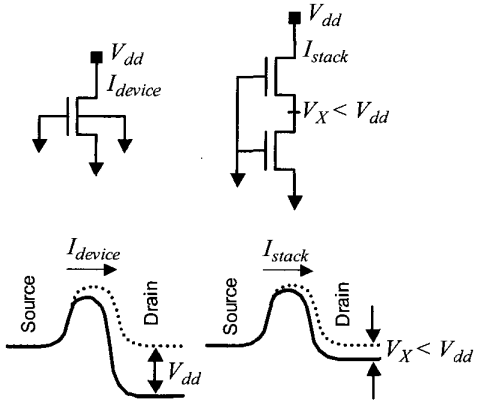

Fig. 1 Leakage current difference between a single *off* device and a stack of two *off* devices. As illustrated by the energy band diagram, the barrier height is modulated to be higher for the two-stack due to smaller drain-to-source voltage resulting in reduced leakage.

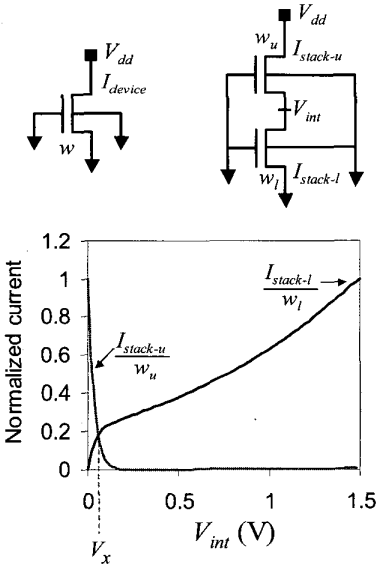

Fig. 3 Load line analysis showing the leakage reduction in a two-stack.

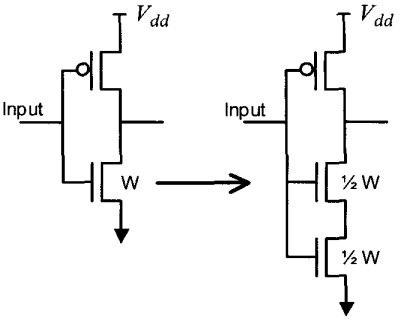

Fig. 2 Trade-off between standby leakage and performance by forcing a two-stack under iso-input load. An NMOS two-stack will reduce leakage when input stays at logic "0"

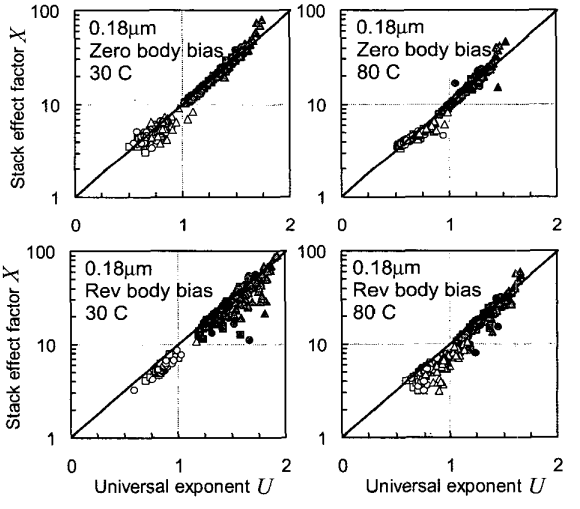

Fig. 4 Measurement results showing the relationship between stack effect factor X for a two-stack to the universal exponent U. Lines indicate the relationship as per the analytical model and symbols are from measurement results. White symbols are for nominal channel devices and gray symbols are for devices smaller than the nominal channel length. Triangle, circle, and square symbols are for V_{dd} of 1.5, 1.2, and 1.1 V respectively. Zero body bias is when the body-to-source diode of the device closet to the power supply is zero biased and reverse body bias is when the diode is reverse biased by 0.5 V.

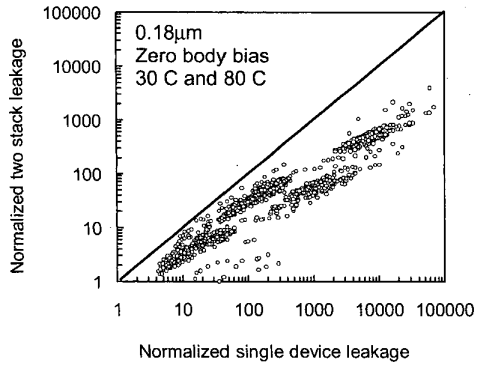

Fig. 5 Measurement results indicate a slower rate of increase in leakage of two-stack compared to that of a single device.

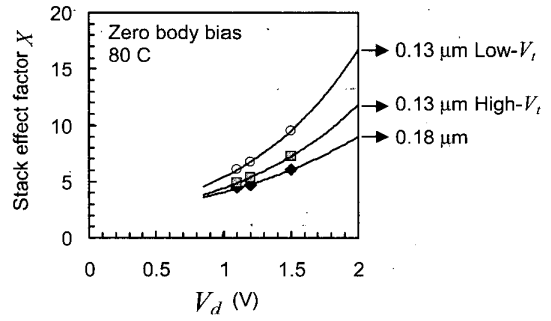

Fig. 6 Device measurement results showing stack effect factor across two technology generations. The increase in stack effect factor is attributed to worsening of short channel effect, λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13μm technology generation is attributed to the same reason. Lines are from analytical model and symbols are from measurement.

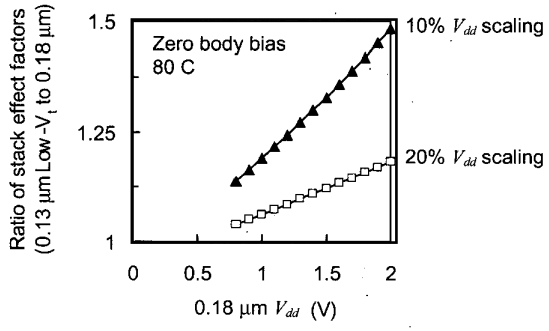

Fig. 7 Device measurement results indicating the scaling of stack effect factor from 0.18μm to 0.13μm low-V_t under different V_{dd} scaling conditions. The low-V_t device will dominate leakage in 0.13μm technology, so the comparison is made with the low-V_t device.

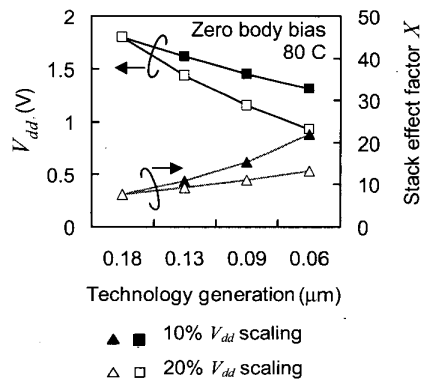

Fig. 8 Prediction in the scaling of stack effect factor for two V_{dd} scaling scenarios. V_{dd} for 0.18μm is assumed to be 1.8 V.

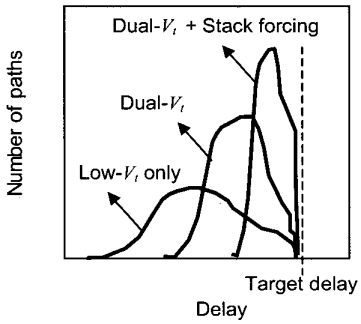

Fig. 9 Stack forcing and dual-V_t can reduce leakage of gates in paths that are faster than required.

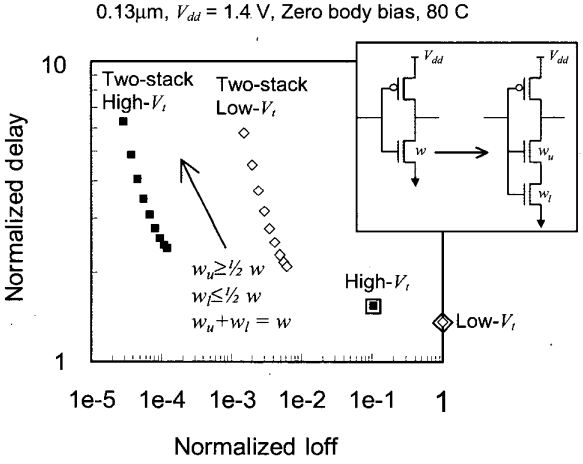

Fig. 10 Simulation result showing the delay-leakage trade-off that can be achieved by stack forcing technique under iso-input load conditions. Iso-input load is achieved by making the gate area after stack forcing identical to before stack forcing. Several such conditions are possible, which enhances delay-leakage trade-off possible by stack forcing. The two-stack condition for a given V_t with the least delay is for $w_u = w_l = ½w$. This trade-off can be used with or without high-V_t transistors.

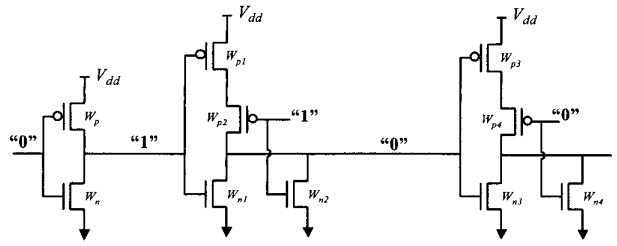

Fig. 11(a) A sample path where natural stack is used to reduce standby leakage by applying a predetermined vector during standby. No delay penalty is incurred with this technique.

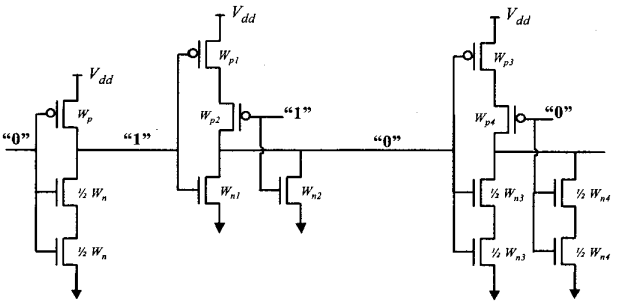

Fig. 11(b) Using stack-forcing technique the number of logic gates in stack mode can be increased. This will enable further leakage reduction in standby mode. Increase in delay under normal mode of operation will be incurred.

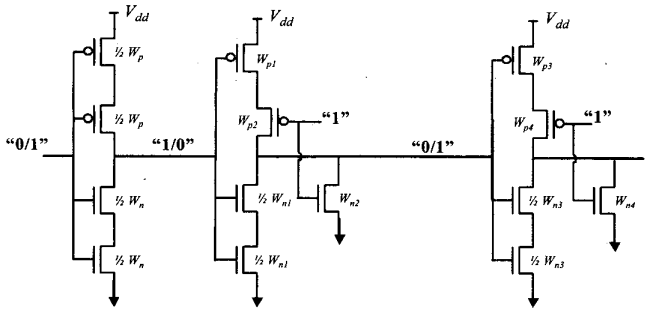

Fig. 11(c) If a gate can have its input as either "0" or "1" and still force stack effect then that gate will have reduced active leakage. The more the number of inputs that can be either "0" or "1" the higher the probability that stack effect will reduce active leakage.

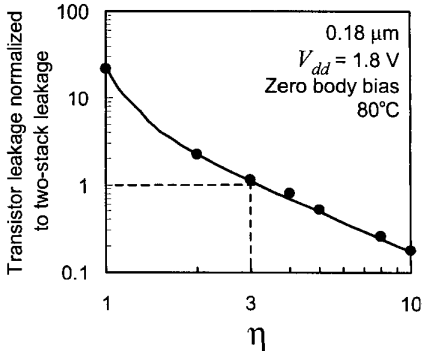

Fig. 12 Comparing device leakage reduction due to channel length increase with two-stack leakage. The channel length is given by η x 0.18 µm. Stack leakage is a two stack of devices with η=1 and w_u=w_l=½w. Leakage numbers are obtained from simulation under iso-input load.

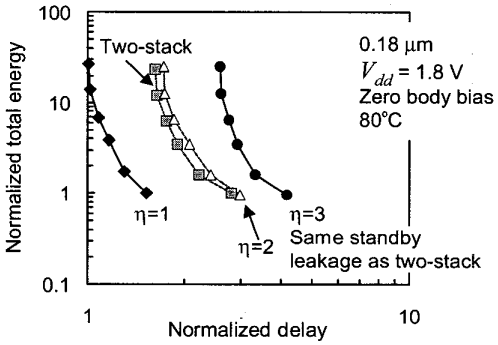

Fig. 13 Energy-delay trade-off of inverter under different configurations with fan-out of 1 and iso-input load. The simulation based comparison clearly shows that the two-stack configuration's delay is less than increasing channel length, especially when compared to iso-standby leakage (η=3) configuration.

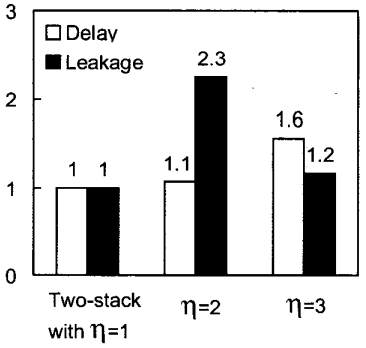

Fig. 14 Summary of delay-leakage trade-off comparison between two-stack and channel length increase.