A New Threshold Voltage Assignment Scheme for Runtime Leakage Reduction in On-Chip Repeaters

Saumil Shah, Kanak Agarwal, Dennis Sylvester
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
E-mail: {saumil, agarwalk, dmcs}@umich.edu

Abstract – High performance digital circuits require long bus lines to operate at very high frequencies, necessitating a large number of repeaters to be inserted along these lines. Power consumed by repeaters, particularly that contributed by subthreshold leakage, is becoming a major consideration in digital design. We compare several threshold voltage assignment schemes to reduce *runtime* leakage power in buffers. We explore trade-offs between dynamic and static power by selectively mixing high and low Vt devices within a pull-up or pull-down network. We propose an activity-dependent hybrid Vt assignment scheme that can be applied across a bus. These configurations are shown to reduce total power by up to 38% and runtime leakage by up to 48%, with negligible design or area overhead.

1. INTRODUCTION

On-chip buses in modern high-frequency processors and ASICs are required to operate at very high speeds to reduce growing communication latencies. To meet these constraints, a large number of repeaters need to be inserted on the buses. It has been observed that, as opposed to the normal scaling of 0.7x per generation, critical wire lengths on buses scale by 0.57x [1]. The number of repeaters therefore increases rapidly across process generations. Furthermore, the need for high throughput and low latency communication calls for the use of aggressively sized devices, making them expensive in terms of power. It was reported that the energy required to transfer 32 bits over a distance of 1 cm in a modern microprocessor is 20X larger than the energy required to perform a 32-bit arithmetic operation [2] and this factor will continue to grow with scaling.

Also due to tight timing constraints low-Vt devices are frequently used in repeaters, leading to enormous subthreshold leakage currents that also increase by 3-5X per technology generation [3]. Static power is expected to grow to 40% of total microprocessor power in the 90nm process technology [4]. Taking these effects into account, repeaters are expected to consume over half of the global wire power dissipation, much of which will be leakage [5]. These trends have led power, especially that due to subthreshold leakage, to become a first-class design consideration in global interconnect planning.

There has been work on using now-standard dual-Vt processes to reduce leakage in the standby state [6], but this alone is not expected to be a feasible solution to the leakage problem. In particular, runtime leakage is an issue that needs to be addressed concurrently with standby power. In general the problem of runtime leakage is much more difficult since any relevant design techniques cannot incur any delay penalties, as opposed to standby mode techniques that inherently rely on periods of circuit inactivity where delay is inconsequential. Most known leakage reduction techniques, including adaptive body biasing (ABB) [7], are expensive in terms of dynamic power or latency, or they incur significant design/manufacturing overhead. In this paper we explore threshold voltage assignment schemes that reduce runtime leakage of repeaters inserted in on-chip buses. The techniques discussed in this paper do not incur any significant design or manufacturing costs and provide considerable savings in both static and total power. The method relies on the availability of a dual-Vt process which is widely used in current process technologies.

The remainder of this paper is organized as follows. Section 2 motivates the need for new Vt assignment schemes and discusses a previously proposed configuration. Section 3 provides a basic theoretical analysis and introduces different configurations. We also discuss the design and manufacturing overheads of these techniques in this section. Section 4 details the experimental setup and discusses the obtained results. Finally, Section 5 concludes the paper.

2. MOTIVATION AND OVERVIEW OF PREVIOUS CONFIGURATIONS

2.1 Motivation

From a purely dynamic energy point of view, low Vt (LVt) repeaters are superior to high Vt (HVt) repeaters, since they can achieve a given delay constraint with smaller device sizes. At high activity factors, dynamic energy forms a large proportion of total power, which makes the use of LVt repeaters desirable. On the other hand, LVt repeaters exhibit very high subthreshold leakage; in the case of low switching activity this enhanced leakage dominates, causing the total power consumption to rise for LVt repeaters relative to HVt. Unfortunately, given stringent performance constraints HVt repeaters cannot meet the timing requirements within reasonable constraints on device sizing. In general, even in cases where HVt repeaters can meet looser timing requirements, they will suffer a delay penalty of roughly 12-15% compared to an LVt configuration at the same size due to their reduced drain current. This indicates that, for a given delay, HVt devices must be considerably larger than LVt devices, which acts to increase dynamic power and chip area.

The above qualitative discussion points to an optimization problem whose solution centers on the tradeoff between dynamic power, leakage, and performance. In this paper we develop new repeater configurations that aim to provide better performance and lower area than HVt devices, with only minor dynamic power penalties relative to the LVt case. Depending on the switching activity of the circuit and operating frequency (i.e., timing constraint), this tradeoff will lead to considerable savings in total power. In the next section we discuss a previously proposed repeater configuration that attempts to make this same tradeoff.

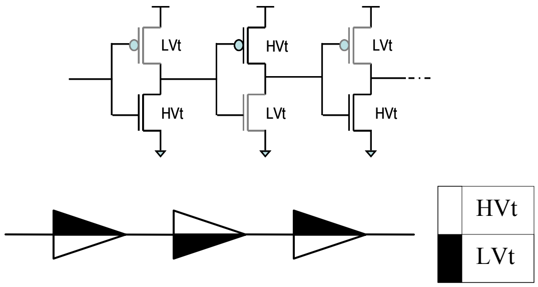

Figure 1. Staggered Vt Configuration

2.2 Staggered Vt Configuration

The Staggered Vth (SVt) configuration proposed in [8] is shown in Figure 1. Here the inverting repeaters are of two types – *RepUp*, where the NMOS is a low Vth Device and the PMOS is a high Vth Device, and *RepDown* where the NMOS is high Vth and the PMOS is low Vth. The energy-delay product (EDP) of this configuration was shown in [8] to besuperior to that of the HVt configuration. It can be readily seen that this configuration has one very low leakage state and one high leakage state. In particular, for the configuration shown in Figure 1 an input of zero gives very low leakage (since all leaking devices are HVt) and a high input yields high leakage (all leaking devices are LVt). Certain buses, such as data buses from caches, may have a high probability of a certain value (typically zeroes), and thus the buffers can be skewed so that they are in the low leakage state the majority of the time. Alternatively [8] focuses on the use of bus encoding techniques to enforce low-leakage states as often as possible. This configuration is also very good for standby leakage, since in standby state, the bus can easily be forced into a low leakage state.

Given a fixed delay constraint and appropriate input state, the SVt configuration can have lower leakage than HVt, since the devices need to be sized smaller to achieve the same operating frequency (shown later). Unfortunately, this scheme is primarily effective in the case that the 0 and 1 state probabilities on the bus can be predicted with reasonable accuracy, which is not always possible. If the data on the bus has a high probability of being in the high leakage state, this scheme yields unacceptably large leakage (close to the LVt case). The fact that devices in SVt must be sized larger than the LVt case for a fixed delay target exacerbates this effect. A final drawback to SVt is that the delay is fully penalized in one transition direction; in particular a rising transition at the input of Figure 1 would propagate through all high Vt devices. This leads to large high-Vt devices for tight timing constraints and ultimately limits how fast the SVt configuration can be operated. Nevertheless, this configuration has certain interesting properties and we discuss its potential applications in Section 3.4.

3. PROPOSED CONFIGURATIONS

3.1 Separate PMOS/NMOS Vt

We first propose a simple modification of the SVt configuration, shown in Figure 2, in which all PMOS devices are low Vt and all NMOS are high Vt. We denote this case as the Separate PMOS/NMOS Vt, or SPNVt. This configuration reduces dynamic energy beyond SVt, since the worst case delay is now evenly distributed between low Vth and high Vth devices.

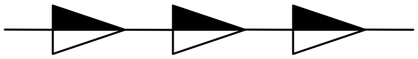

Figure 2. SPNVt Configuration

This limits the up-sizing required for a given delay and thus the EDP is lower than both the HVt and SVt configurations. Also, leakage becomes state-independent and is significantly lower than the LVt configuration, since in either state half the leaking devices are LVt while half are HVt. For typical input state probabilities, this new configuration provides comparable leakage savings (compared to LVt) as the SVt configuration. SPNVt is very easy to design and manufacture, as there is no need for Vt assignment at all. Also, since PMOS and NMOS devices have their threshold voltages set by different ion implant steps, there is no extra manufacturing cost/effort. Note that a similar configuration could be used with high Vt PMOS and low Vt NMOS. However, this approach is sub-optimal since it results in more substantial up-sizing of the weaker PMOS device and degrades the final power savings.

3.2 Mixed Vt Configuration

The SPNVt configuration effectively partitions the total device width associated with a propagating signal between LVt and HVt, providing substantially lower leakage as compared to using all LVt. However, in the SPNVt configuration the width allocation to LVt and HVt is constrained by the fact that all PMOS devices should be made HVt while all NMOS devices should me made LVt. Hence, the SPNVt configuration provides limited control over how total drive strength (width) can be divided between LVt to HVt to obtain maximum savings in power. In this section, we propose a new technique where the ratio of LVt to HVt widths can be controlled in a much more fine-grained manner by mixing low and threshold widths in NMOS/PMOS devices within a repeater. We show that this configuration allows us to obtain the optimal trade-off between all HVt (high dynamic power, low static power) and all LVt (high static power, low dynamic power) configurations.

To illustrate the above claims analytically, we first consider the case where total width (corresponding to dynamic power) of each repeater is fixed. In order to obtain maximum savings in power, we must allot the least possible width to LVt while still meeting our delay requirement. We show that this is possible only when the total LVt width is divided uniformly across all repeaters. This is analogous to the mixed Vt configuration where low and high threshold voltages are mixed within repeater pull-up or pull-down stacks. The analysis compares mixed Vt to the case where LVt-HVt width allocation can be done only at the repeater level (i.e., PMOS/NMOS devices in the repeaters can either be all LVt or HVt but are never mixed).

To a first approximation, the 50% delay of a wire segment of resistance R_w and capacitance C_w driven by a repeater with resistance R_d and fanout capacitance C_L is given by [9]

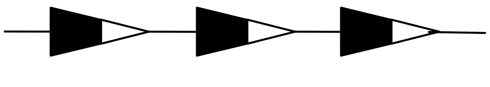

Figure 3. Mixed Vt Configuration

$$D = 0.69 R_d C_L + 0.69 R_d C_w + 0.69 R_w C_L + 0.38 R_w C_w \qquad (1)$$

where R_d can be approximated by $(0.8Vdd)/W*Id_{sat}$ [10]. Id_{sat} is the drain current per micron width of the MOSFET in saturation, obtained from SPICE simulations for the industrial 0.13 µm process used in this paper.

Assume that initially the repeater is HVt. The change in delay when the repeater is converted from HVt to LVt is

$$\Delta D = R_{d(LVt)}(0.69C_L + 0.69C_w) - R_{d(HVt)}(0.69C_L + 0.69C_w) \qquad (2)$$

For this analysis we assume that the fanout capacitance C_L is a function only of width, and not of device threshold voltage. Although this is not strictly true as pointed out in [11], the difference is found to be a marginal 6-8%. This dependence is therefore ignored in the current analysis.

Substituting the expression for R_d in (2), ΔD can be written as

$$\Delta D = 0.552 Vdd (C_L + C_w) \frac{(Id_{sat(HVt)} - Id_{sat(LVt)})}{W * Id_{sat(HVt)} * Id_{sat(LVt)}} \qquad (3)$$

We assume that the initial configuration (all HVt) cannot meet our delay requirement and the total negative slack is S. To meet timing, the number of repeaters that need to be made LVt is

$$n_{LVt} = \frac{S}{\Delta D} \qquad (4)$$

Using (3) and (4), we find that the required LVt width is

$$W_{LVt} = n_{LVt} W = \frac{S * W^2 * Id_{sat(HVt)} * Id_{sat(LVt)}}{0.552 * Vdd * (C_L + C_w)(Id_{sat(HVt)} - Id_{sat(LVt)})} \qquad (5)$$

We now compare this to the distributed approach where the total LVt width is divided equally between all repeaters. Each repeater has a fraction α of its total width allocated to LVt.

Now, the change in delay compared to all HVt is:

$$\Delta D = 0.552 * Vdd (C_L + C_w) \frac{(Id_{sat(HVt)} - Id_{sat(eff)})}{W * Id_{sat(HVt)} * Id_{sat(eff)}} \qquad (6)$$

where $Id_{sat(eff)} = \alpha Id_{sat(LVt)} + (1 - \alpha) Id_{sat(HVt)}$ (7)

The delay constraint is satisfied when the total delay change $n\Delta D$ becomes equal to the negative slack, S.

$$S = n * 0.552 * Vdd (C_L + C_w) \frac{(Id_{sat(HVt)} - Id_{sat(eff)})}{W * Id_{sat(HVt)} * Id_{sat(eff)}} \qquad (8)$$

The total width that needs to be allocated to LVt is now given by

$$W_{LVt} = n\alpha W = \frac{S * W^2 * Id_{sat(HVt)} * Id_{sat(eff)}}{0.552 * Vdd * (C_L + C_w)(Id_{sat(HVt)} - Id_{sat(LVt)})} \qquad (9)$$

Since $I_{dsat(eff)} < I_{dsat(LVt)}$, W_{LVt} as given by (9) is smaller than that computed in (5). Therefore, distributing the LVt width uniformly over all repeaters is more useful than having different Vts for different repeaters. Note that SPNVt is similar to a case where half the repeaters are HVt and half are LVt, and is, therefore, covered in this analysis.

The preceding analysis motivates a configuration in which each repeater is divided into multiple (we assume ten in this work) parallel fingers. Note that this is reflective of actual layout practices for very wide gates in order to maintain reasonable cell aspect ratios, reduce junction capacitances, and limit gate resistance. Of these fingers, a certain proportion (α, where 0 < α < 1) consists of LVt devices and the remainder (1 − α) use HVt devices. This mixed Vt assignment allows timing to be met with the minimum possible leakage power. This is achieved by seeking the smallest α that achieves the required speed. Note that if the timing constraint is very loose (i.e., HVt can readily meet it) or very tight (i.e., only LVt can meet) the required α converges to 0 or 1 respectively, making this a general technique. The next section provides a brief and intuitive theoretical analysis that supports the efficiency of the mixed Vt configuration and explores the various tradeoffs involved in selecting α.

3.3 Theoretical Analysis

We assume that the P/N devices have been sized for equal drain saturation currents, with W_P/W_N equal to 2. C_L is given by $W*C_{ox}*L$. In this analysis we include the effect of Vt on C_L in order to obtain a more accurate relation among power, activity factor, and α. The average values of C_{ox} for both device types are obtained using SPICE simulations.

The complete equation for C_L is now

$$C_L = W * L * C_{ox(eff)} \qquad (10)$$

where

$$C_{ox(eff)} = C_{ox(eff)P} + 0.5 * C_{ox(eff)N} \qquad (11)$$

$$C_{ox(eff)P/N} = \alpha C_{ox(LVt)P/N} + (1 - \alpha)(C_{ox(HVt)P/N}) \qquad (12)$$

From the equations given above, we can obtain an expression for the PMOS width W in terms of the given delay, as well as all other constant parameters by using equations discussed in the previous section.

$$W = \frac{(D - K_1 - K_4) - \sqrt{(D - K_1 - K_4)^2 - 4 * K_2 * K_3}}{2 * K_3} \qquad (13)$$

In this equation, the values of K_1, K_2, K_3, and K_4 are:

$$K_1 = \frac{0.552 * Vdd * C_{ox(eff)} * L}{\alpha * Id_{sat(LVt)} + (1 - \alpha) * Id_{sat(HVt)}} \qquad (14)$$

$$K_2 = \frac{0.552 * Vdd * C_w}{\alpha * Id_{sat(Lvt)} + (1 - \alpha) * Id_{sat(Hvt)}} \qquad (15)$$

$$K_3 = 0.69 R_w C_{oxeff} L \qquad (16)$$

$$K_4 = 0.38 R_w C_w \qquad (17)$$

We also express the dynamic and static power in terms of the operating frequency *f*, activity factor *A*, and device size W.

$$P_{dyn} = A \cdot (C_L + C_w) \cdot Vdd^2 \cdot f \qquad (18)$$

$$P_{stat} = Vdd * W * I_{off} \qquad (19)$$

Since the subthreshold leakage (or off) current I_{off} is different for LVt and HVt devices, P_{stat} can be written as

$$P_{stat} = Vdd * W * (\alpha * I_{off(LVt)} + (1 - \alpha) * I_{off(HVt)}) \qquad (20)$$

$$P_{total} = P_{stat} + P_{dyn} \qquad (21)$$

Here we again assume that the I_{off} values of P and N are equal when the P:N sizing ratio is 2.

To a first order, P_{stat} increases linearly with α. This occurs since in the mixed Vt case static power is a linear combination of $I_{off(LVt)}$ and $I_{off(HVt)}$, as shown in (20). Equation 20 also contains a W term, causing additional dependency of P_{stat} on α since a smaller α will require larger device sizes to meet a given timing constraint. Typically, the dependence of P_{stat} on I_{off} is much stronger than the W dependency, causing P_{stat} to increase monotonically with α. On the other hand, P_{dyn} varies inversely with α due to the aforementioned need for larger device sizes as α (LVt fraction) reduces. The sum of these two power terms (Equation 21) is also nearly linear, and the sign of the slope depends on the value of activity factor A.

$$A < \frac{W_{LVt} I_{off(LVt)} - W_{HVt} I_{off(HVt)}}{Vdd * f * L * (W_{HVt} C_{ox(HVt)} - W_{LVt} C_{ox(LVt)})} \quad (22)$$

It can be shown that when the inequality in (22) is satisfied, power in the all HVt case is less than that in the all LVt case, wheareas for larger values of A, the converse is true. Using values obtained from device characterizations in SPICE and an operating frequency of 2 GHz, this crossover point evaluates to A=0.23 for our process. Examining the switching behavior on address buses in several architecture-level benchmarks applications, the average activity was found to be 0.13. Note that as I_{off} increases in future technology generations, the crossover point will move closer to the maximum value of 0.5. This shows that it is clearly beneficial to operate at lower values of α. In Section 4.2, we will show that simulation results exhibit a similar trend, although the crossover point does not match exactly due to the approximations made. The equation confirms the intuitive explanation that for low activity factors, allotting a large proportion of the width to HVt (corresponding to a low value of α) provides lower total power, whereas for high activity increasing the LVt width (increasing α) reduces power consumption. Choosing an intermediate alpha, we obtain acceptable power values, along with the ability to operate the bus at high frequencies. Even if the timing constraint can be met by HVt repeaters, it is clear that the mixed Vt configuration always provides better performance than the worst case (where HVt buses are operated at high activities or LVt buses at low activities). Mixed Vt avoids unacceptably large power numbers when the activities of the bitlines are unknown, and also achieves optimal energy-delay performance for a large range of activity factors.

3.4 Hybrid Configuration

In certain situations, the activity of some lines in a bus may be known at design time. For example, an Alpha-architecture based processor has a quadword access. In this case, the least significant two bits of the address bus are nearly always zero. Furthermore, some 64-bit address buses have very low activity on the most significant 32 bits. Also, as previously discussed, data output buses from on-chip caches store primarily zeroes. In such cases, it can be highly beneficial to use a hybrid configuration, where the lines known to be skewed towards a particular state use the SVt scheme and the remaining lines use the mixed Vt scheme. The buffers must be designed carefully to ensure that the most probable state is the low leakage state. When correctly designed, using the hybrid configuration can yield greater power savings than the mixed Vt configuration by itself, depending on the number of bits customized, and the disparity between the 0 and 1 state probabilities.

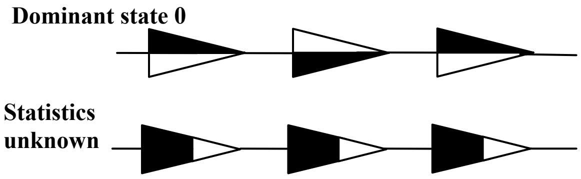

Figure 4. Hybrid Configuration

3.5 Design/Manufacturing Overheads

In this section we consider the design, manufacturing, and area overheads for the above Vt assignment schemes.

As mentioned earlier, the SPNVt scheme incurs no manufacturing or area overhead, as the PMOS and NMOS devices are fabricated in separate wells. The SVt configuration also uses different Vt in the pull-up and pull-down stack and hence does not incur any manufacturing or area overheads. The mixed Vt scheme relies on different threshold voltages for parallel-connected devices where there is a relatively large spacing due to the intervening contact. Due to the presence of this large spacing, the threshold adjust ion implantation tolerance is adhered to without a corresponding increase in poly spacing [12]. This is in contrast to the case of uncontacted poly pitch, which must be increased substantially to tolerate different threshold voltages in adjacent (series-connected) devices. Thus, the mixed Vt scheme is highly manufacturable and does not incur any discernible area penalty. At the same time, the mixed Vt scheme combines properties of high and low Vt devices in a dual-Vt process. Thus it provides flexibility to the designer, with no manufacturing costs and no silicon area costs.

The increase in library complexity of each of these configurations is also small. The SPNVt configuration adds another cell to the library for each repeater size, and SVt requires two new cells to be added. For the mixed Vt configuration, we need to add one new cell type to the library (for a given buffer size). This added cell type complements the pre-existing high-Vth and low-Vth flavors of the buffer. However, if more than one α value is to be made available the library size would grow accordingly. This represents a fundamental tradeoff between library complexity and performance; we will see in the next section that a single α ratio can provide good results over a range of benchmark applications.

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

We use an industrial 0.13 μm CMOS process at 1.2V supply. All simulations are run at a temperature of 105C. An 8mm line is used with repeaters inserted every 800μm. The parasitic interconnect resistances and capacitances were extracted based on the Berkeley Predictive Technology Model [13].

The setup focuses on a 3-bit portion of a bus. Delay and power measurements are carried out under the conditions of worst-case crosstalk – with the middle bit line switching in one direction and both neighbors switching in the opposite direction. The widths of the devices used in the repeaters are swept over a wide range. Energy-delay curves are plotted and a suitable delay point is chosen from the curves to reflect a practical operating environment. The power numbers obtained in our simulations are used along with statistics obtained from various benchmark application traces. These traces are taken from the address bus of an Alpha-architecture based processor [14]. Although the data is for a 64-bit address bus, we use only the lower 32 bits, as the upper 32 bits show almost no activity. If the upper 32 bits are also considered, they can be assigned to a low leakage configuration (SVt, biased to the low leakage state). In this case, the power savings obtained are greatly increased as discussed in Section 3.4.

4.2 Results

Figure 5 plots dynamic energy vs. delay for the various configurations discussed using the setup of Section 4.1. We see that LVt has the lowest dynamic energy at a given delay; HVt exhibits the worst behavior, while all other configurations lie between these two extremes. The mixed Vt case clearly provides the closest performance to LVt, while the SVt configuration is similar to HVt. These plots show the dynamic power tradeoff involved, which can be recovered by leakage savings. Table 1 shows the total power for different α values over various switching activities. Figure 6 is a graphical representation of the data provided in Table 1. Clearly the data confirms to the trend predicted by equation (22). We integrate the data with the switching activity factors and 0/1 state probabilities of various bus lines obtained from the application traces. The delay constraint is such that it cannot be met by all HVt, given a reasonable limit on device width.[1] This, as stated earlier, is commonly the case in high performance processors. The exact delay constraint is set midway between the fastest possible LVt and HVt configurations given the size constraint. In this case, α of less than 0.3 is not able to meet timing and thus results for these cases are not shown. Table 2 shows the normalized total power of various configurations over all benchmarks. Figure 7 compares the total power consumption of the best case (Hybrid Vt, where the bottom two bits are SVt due to dominant 0 states and top 30 bits are mixed Vt) to the LVt configuration for all the benchmarks.

The figure also shows the individual contributions of leakage and dynamic power to total power. It is seen that there is considerable reduction in total power for all benchmarks. The maximum reduction is seen when the switching activity is very low, in the case of the CRAFTY benchmark. In this case, a 38% reduction is achieved, as shown in Table 2 and Figure 7. The average total power reduction is 11% across all applications.

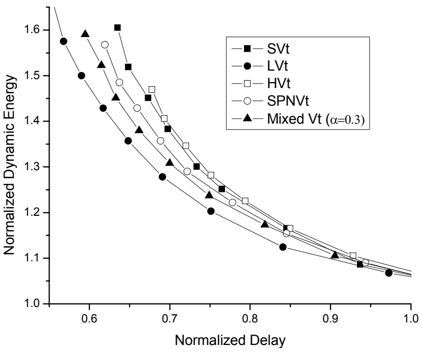

Figure 5. Dynamic energy delay curves for different configurations.

Table 1. Total power consumption (normalized) of different α values for different switching activity factors.

Activity →	0.05	0.1	0.2	0.5
Alpha				
0.3	0.811	0.887	0.942	1.010
0.4	0.842	0.906	0.952	1.008
0.5	0.872	0.923	0.960	1.006
0.6	0.899	0.940	0.969	1.005
0.7	0.926	0.956	0.977	1.003
0.8	0.952	0.971	0.985	1.002
0.9	0.976	0.986	0.992	1.001
1 (LVt)	1.000	1.000	1.000	1.000

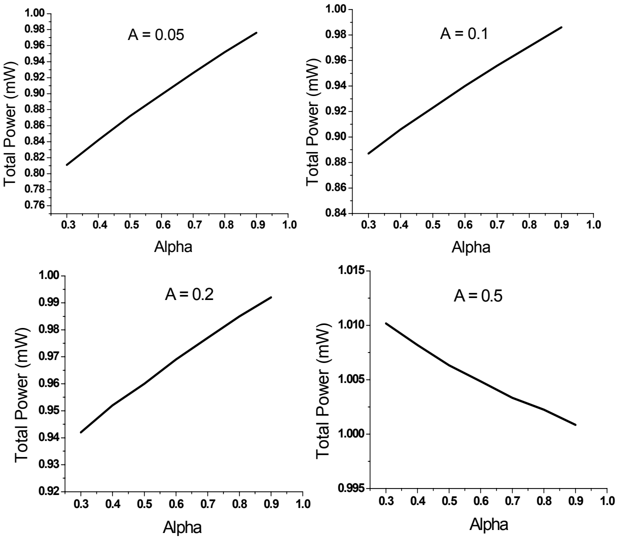

Figure 6. Sensitivity of total power to α for various switching activity factors. Note different y-axis scales.

Table 2. Total normalized power consumption of the various configurations for different benchmarks.

Config →	LVt	SVt	SPNVt	α = 0.3	Hybrid
Benchmark					
BZIP	1	0.955	0.930	0.890	0.894
CRAFTY	1	0.713	0.701	0.627	0.613
EON	1	0.936	0.926	0.886	0.881
GAP	1	0.981	0.971	0.937	0.937
GCC	1	0.995	0.999	0.946	0.943
GZIP	1	0.961	0.964	0.929	0.927
MCF	1	0.891	0.913	0.871	0.865
PARSER	1	0.951	0.952	0.916	0.913
TWOLF	1	0.953	0.955	0.919	0.917
VORTEX	1	0.948	0.938	0.900	0.895
VPR	1	0.967	0.964	0.930	0.927

[1] This limit is set to a device width of 30μm in this case.

Table 3. Normalized runtime leakage of various configurations under study

Config	LVt	α=0.3	SPNVt	SVt(LL)	SVt(HL)
Leakage	1	0.521	0.763	0.243	1.025

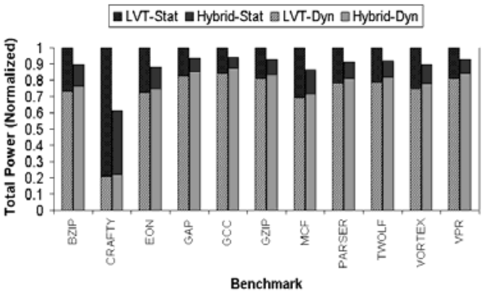

Figure 7. Comparison of hybrid scheme with LVt for all benchmarks.

Table 3 shows the normalized runtime leakage power consumption of the different configurations. Although, the lowest possible leakage is in the case of SVt (Low Leakage State), the total leakage of SVt is heavily dependent on the probabilities of the bus holding a 0 or a 1. Except in certain cases (as discussed in Section 3.4) this is not readily determined at design time. Therefore, on average the SVt configuration does not give reductions that are as significant as Mixed Vt given actual data.[2] Also, the plots in Figure 5 show that the SVt configuration has higher dynamic energy compared to mixed Vt. This is due to the fact that the critical path of the SVt configuration is through all high Vt devices. It is clear that, although SVt by itself should not be used due to uncertainties in the 0/1 probabilities, the low leakage state is very attractive if one particular state can be known to be dominant. The results also show that the hybrid scheme gives improvement over using any of the configurations by itself. Recall that this improvement is obtained by placing SVt buffers on just two of the 32 lines. The possibility of such customizations leads to interesting asymmetric Vt assignment problems, which can have compelling and effective solutions.

5. CONCLUSIONS AND FUTURE DIRECTIONS

To address the issue of increasing runtime leakage power consumption of repeaters, several different Vt assignment schemes have been analyzed. The static/dynamic power tradeoffs in each of the schemes have been explored. Selective Vt assignment techniques were proposed that give considerable reductions in runtime leakage with minimal dynamic power overhead. Depending on the requirements, the designer can optimize static/dynamic power trade-offs simply by changing the proportion of high and low threshold devices.

This technique has the potential for wide application in future process technologies. With process scaling, the proportion of leakage further increases, thus increasing leakage savings for a larger range of activity factors. We also note that, with scaling, a fixed 100mV Vt offset in a dual Vt process causes a growing disparity in the delay characteristics of low and high Vt repeaters. In cases where the timing and/or leakage disparity between HVt and LVt is very large, using either all HVt or all LVt becomes increasingly sub-optimal from the delay and power perspectives respectively. The mixed Vt scheme provides the designer with an extra degree of freedom in the Vt-assignment problem, and can therefore lead to more optimal solutions across a range of applications. The mixed Vt scheme is also less susceptible to process variations as compared to all LVt, since variability in low Vt devices is known to be appreciable [15]. Finally, mixed Vt has greater potential when considering emerging devices such as FinFETs [16]. These devices can only be made at a single fixed width and thus will require multiple parallel fingers even for moderate device widths. This lends itself naturally to the application of the mixed Vt scheme.

In summary, the newly proposed mixed Vt assignment scheme achieves reductions of up to 48% in runtime leakage and 38% in total power with an average total power reduction of 11%. The proposed mixed Vt configuration requires very little design effort and has negligible area or manufacturing overhead.

[2] For example, if input states are equally likely the average leakage in SVt becomes 0.634.

REFERENCES

[1] P. Saxena, *et. al*., "The scaling challenge: can correct-by-construction design help?", *Intl. Symp. on Physical Design*, pp. 51-58, 2003.

[2] W.J. Dally, "Computer architecture is all about interconnect", *High-Perf. Comp. Architecture,* panel discussion, 2002.

[3] S. Borkar. "Design challenges of technology scaling", *IEEE Micro,* pp. 23-29, July/Aug 1999.

[4] J. Kao, S. Narendra, and A. Chandrakasan. "Subthreshold leakage modeling and reduction techniques", *Intl. Conf. on Computer-Aided Design,* pp. 141-148, 2002.

[5] V. Kapur, G. Chandra, and K.C. Saraswat, "Power estimation in global interconnects and its reduction using a novel repeater optimization methodology", *DAC,* pp. 461-466, 2002.

[6] S. Mutoh *et. al*., "A 1-V power supply high-speed digital circuit technology with multithreshold voltage CMOS", *IEEE Journal of Solid State Circuits*, pp. 847-854, Aug. 1995.

[7] H.C. Wan *et al*., "Channel doping engineering of MOSFET with adaptable threshold voltage using body effect for low voltage and low power applications", *Intl. Symp. VLSI Technology, Systems, and Applications*, pp. 159–163, 1995.

[8] H. Deogun, *et. al*., "Leakage and Crosstalk-aware Bus Encoding for Total Power Reduction", *DAC ,* pp. 779-782, 2002.

[9] H. B. Bakoglu, "*Circuits, Interconnections and Packaging for VLSI*", Addison-Wesley, 1990.

[10] J. Rabaey, A. Chandrakasan and B. Nikolic, "*Digital Integrated Circuits : A Design Perspective*", Prentice-Hall, 2002.

[11] S. Sirichotiyakul, *et. al*., "Duet: An accurate leakage estimation and optimization tool for dual-Vt circuits", *IEEE Trans on VLSI Systems*, pp. 79-90, April 2002.

[12] K. Bernstein, IBM TJ Watson Research Center, personal communication.

[13] Berkeley Predictive Technology Model (BPTM), http://www-device.eecs.berkeley.edu/~ptm/

[14] T. Austin, E. Larson and D. Ernst, "Simplescalar: an infrastructure for computer system modeling", *IEEE Computer,* pp. 59-67, Feb 2002.

[15] R. Puri, IBM TJ Watson Research Center, personal communication.

[16] X. Huang, *et. al*., "Sub 50-nm FinFET: PMOS", *IEDM Technical Digest*, pp. 67-70, 1999.