

Logic Design Considerations for 0.5-Volt CMOS *

K. Joseph Hass Jack Venbrux Prakash Bhatia
NASA Institute of Advanced Microelectronics
University of New Mexico
jhass@mrc.unm.edu jvenbrux@mrc.unm.edu pbhatia@mrc.unm.edu

Abstract

As the operating supply voltage for commercial CMOS devices falls below 2 V, research activities are underway to develop CMOS integrated circuits that can operate at supply voltages well under 1 V. Although dramatic power reductions can be achieved using low supply voltages in high performance applications, the increased subthreshold leakage that results when transistor threshold voltages are lowered can render some conventional logic circuit styles unusable. Furthermore, some low voltage circuits are not robust when faced with normal variations in threshold voltage. This paper examines the design considerations for logic and memory circuits in very low voltage CMOS, and compares simulated behavior with measurements of fabricated test circuits. These circuit examples were chosen because they illustrate the unique design challenges of low voltage CMOS.

1. Introduction

Driven by the need to reduce power consumption and maintain high reliability in leading edge integrated circuits, the nominal operating supply voltage for these devices is falling steadily [3–6, 9]. In order to maintain high switching speed at low supply voltages it is necessary to reduce the transistor threshold voltage, V_T , in proportion. For supply voltages much below 1 V the value of V_T may be just tens or hundreds of millivolts, and the subthreshold leakage current of these transistors becomes significant [2].

However, the power savings achieved by operating at low voltages can be much larger than the power lost to increased static current. The total average power consumption of a CMOS circuit can be expressed as $P_{TOTAL} = P_{STATIC} + P_{DYNAMIC}$ if we focus on these two primary components. To first order, P_{STATIC} is proportional to V_{DD} and $P_{DYNAMIC}$ is proportional to $V_{DD}^2 \times C_{LOAD} \times F_{CLOCK}$, where V_{DD} is the supply voltage, C_{LOAD} is the average capacitance that must be switched in each clock cycle, and F_{CLOCK} is the operating clock frequency. Because of the V_{DD}^2 nature of dynamic power, dramatic gains in overall power consumption can be made by reducing the supply voltage. If we simultaneously decrease V_T to maintain performance then the static power component will increase, but not as quickly as dynamic power decreases. The minimum average power operating point for any circuit occurs when the supply voltage is reduced to the point that the dynamic power and static power components are roughly equal.

However, there are additional engineering challenges that appear when attempting to use transistors with very small V_T values [8]. The normal variation of V_T during manufacturing, and with

*This research was supported by NASA under Space Engineering Research Grants NAG5-8392 and NAG5-7360.

environmental factors like temperature, can be a significant fraction of the desired V_T value. This variation makes it very difficult to fix V_T at a level that will result in robust circuit operation. One approach to solving this problem is to abandon the concept of V_T as a constant parameter. Instead, V_T is treated as an operating parameter that can be adjusted in real time to compensate for manufacturing and environmental effects [7, 10, 11]. To accomplish this the transistors are fabricated with a native, or intrinsic, threshold voltage that is very near zero. Bias voltages are then applied to the transistor body, through the substrate or well, to raise the threshold to the desired level. We refer to the PMOS body bias voltage as *PBIAS*, which is a positive voltage from V_{DD} to the N-well. Similarly, the *NBIAS* is the magnitude of the voltage that drives the substrate (or P-well) to a level below V_{SS} . This biasing technique can compensate for global manufacturing variation and environmental effects, leaving the local variations in V_T across a single die as the limiting factor in reducing the supply voltage.

Figure 1 shows relative NMOS drain current as a function of gate voltage for several different NBIAS voltage settings, illustrating how the transistor's behavior can be tuned during operation. The bias voltages are completely independent of the supply voltage, and as they form reverse biased PN junctions their current demands are quite low.

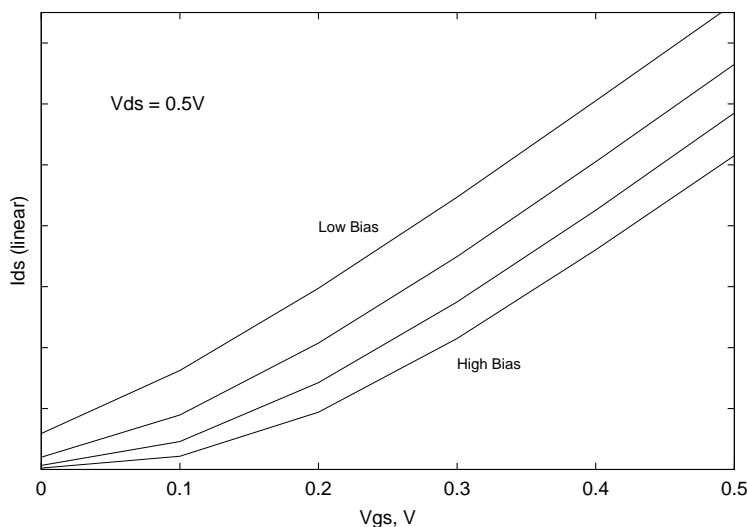


Figure 1. Typical NMOS I/V Curves

Note that the maximum saturation current for this transistor increases by about 50% with low bias voltages, while the subthreshold leakage at low bias is several times higher than the leakage with a high bias. An important operating parameter for these circuits is the ratio of saturation current to leakage current, referred to as the I_{ON}/I_{OFF} ratio. At low bias voltages this ratio is low, but the net drive current is at its peak. On the other hand, high bias voltages greatly reduce subthreshold leakage but the lower drive current reduces performance.

We have fabricated and demonstrated complex microcircuits operating with a supply voltage of only 0.5 V using this technology [1]. Some of the CMOS logic styles that have been used successfully in conventional processes were found to work poorly, if at all, in the low voltage process. Examining these circuits provides a general insight into the appropriate design styles for logic and memory that are intended for operation at a very low supply voltage.

2. Static latches and flip flops

A number of flip flop circuits were fabricated on the first wafer lots of the low voltage process, before accurate SPICE models were available. Examining two of these circuits illustrates a common problem with subthreshold leakage in ratioed logic. The circuit on the left side of Figure 2 is a simple master latch from one of the experimental flip flops. The latch is intended for static operation and has been used successfully in conventional CMOS processes. The feedback path in the latch uses very weak transistors, MN1 and MP1. These devices have a W/L ratio of only 0.16 and are easily over driven by the input pass transistor (W/L = 4.57) to change the state of the latch. In conventional CMOS the weak feedback inverter is more than adequate to supply any leakage current through transistor MN0 at the feedback node and maintain a suitable voltage level. Unfortunately, in a very low voltage process the leakage current through transistor MN0 is a large fraction of the drive current available from MN1 or MP1. The ratio of W/L between the pass transistor and the feedback devices is about 29, and this circuit will fail if I_{ON}/I_{OFF} is less than several times that value.

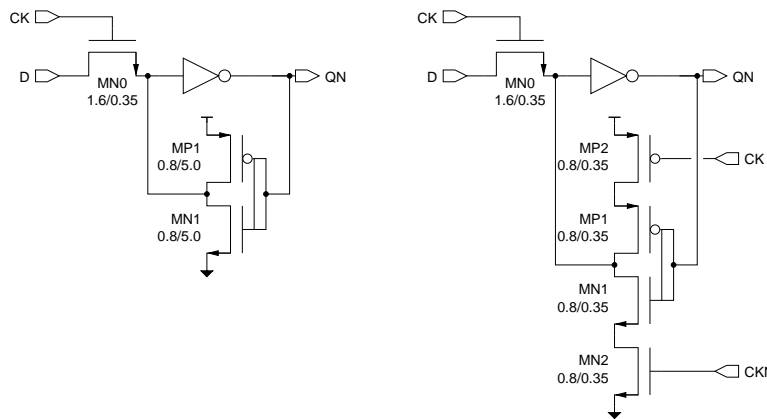


Figure 2. Schematic of Latch Circuits

The latch circuit shown on the right side of Figure 2 is a simple modification to the original latch that provides much better operation. The weak inverter has been replaced with a clocked inverter that has considerably higher drive. The feedback devices now have a W/L ratio half as large as that of the input pass transistor, and they can maintain the desired logic level at the feedback node. The disadvantage of this circuit is that the clock signal must be routed to the added transistors and the capacitive load on the clock nets will be increased.

The operation of the improved latch circuit was evaluated over a range of back bias conditions, as shown in Figure 3. The left shmoo plot indicates those combinations of NBIAS and PBIAS that result in correct function of a sample 32-bit shift register using this latch circuit. The shmoo plot on the right presents SPICE simulations of the flip flop from the shift register, using SPICE transistor models extracted from the same wafer lot. In both cases the supply voltage was 0.5 V and the clock frequency was 1 MHz. Although the shift registers will operate at much higher frequencies any failure modes due to leakage are exacerbated at low frequencies.

The failure mode for either latch circuit occurs when it is first loaded with a high level. Suppose that after the pass transistor is disabled the D input is driven to a low level. The leakage of the NMOS pass transistor must be supplied by the PMOS transistors in the feedback circuit, which have relatively lower drain currents than the NMOS transistors because of the lower mobility of

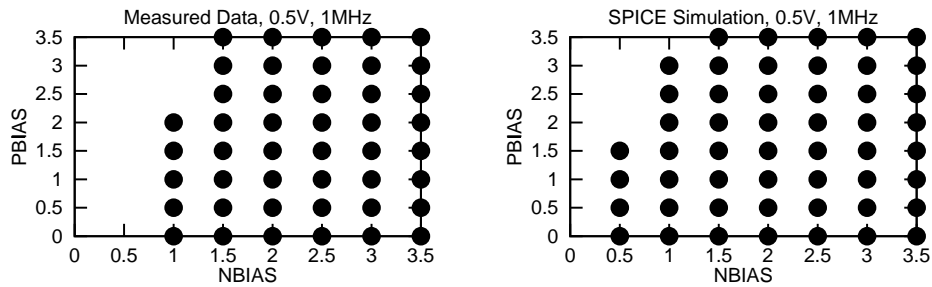


Figure 3. Measured and Simulated Shmoo Plots for Improved Latch Circuit

holes. The latch fails when a valid ‘1’ logic level cannot be maintained by the feedback circuit.

SPICE predicts correctly that the improved latch circuit will not function for an NBIAS of 0.0 V, regardless of the PBIAS voltage. Under these conditions the leakage of the input pass transistor is simply too large for the feedback circuit to supply. At somewhat higher values of NBIAS the latch will function only with lower values of PBIAS because the PMOS transistors in the feedback path must have fairly low values of V_T to provide sufficient current. For NBIAS levels of 1.5 V and above the NMOS leakage is reduced to a low enough level that the PMOS feedback devices can supply this current at any PBIAS value.

Measured data from 32-bit shift registers built using these two latch variations is shown in Figure 4, and provides a macroscopic view of their behavior. This data was collected with a supply voltage of 0.5 V and a clock frequency of 1 MHz. The top trace in this figure is the input data waveform, consisting of a “010110” pattern surrounded by zeros. The second trace in the figure shows the correct shift register output as generated from a shift register using the improved latch. The third and fourth traces show the incorrect behavior of the original latch circuit. With a low bias voltage the leakage of the input pass transistor overwhelms the weak feedback inverter and the input data flows asynchronously to the register output. When the bias voltage is higher the shift register begins to show some signs of synchronous operation but the flip flops still do not work properly.

3. Multiplexer examples

In the previous section, latches were discussed which illustrated some of the design concerns using very low voltage. In this section, three different multiplexer circuits will be discussed that will provide additional insights as to the sensitivities of circuit topology to differences in delay and power given low and high thresholds. The schematic diagrams for these three circuits are shown in Figure 5. The circuit shown in Figure 5(a) is a multiplexer that is very regular in structure. It is also representative of some types of programmed transistor arrays that insert transistors or short out transistor connections at various points in the array [12]. The second circuit shown in Fig 5(b) is a binary tree structure (BTS) [13]. The leaves of the tree are inputs and the tree reduces down to a single 2 to 1 multiplexer for both the N devices and P devices. These first two multiplexer types use transmissions gates with both N and P devices while the third type of multiplexer uses an NMOS BTS input combined with a PMOS pull-up to restore the degraded ‘1’ logic level that would result at node X. The non-BTS and BTS multiplexer designs will be discussed first followed by the NMOS multiplexer.

The SPICE models were extracted from wafers fabricated using the low voltage process. The intrinsic thresholds of the N and P devices were very low and were increased through the use of

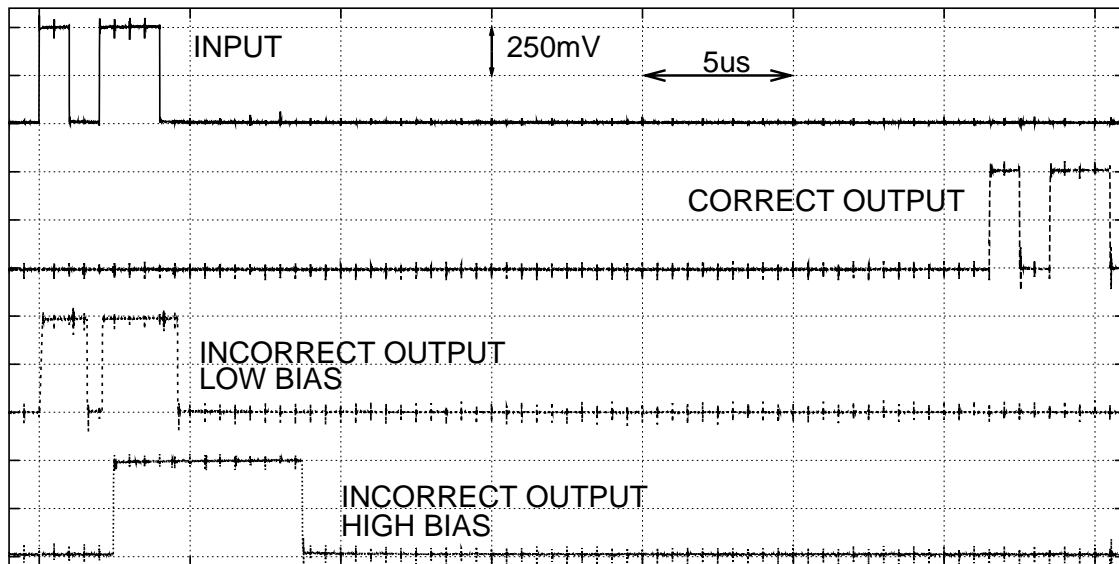


Figure 4. Measured Waveforms for 32-bit Shift Register

back bias voltages as described in Section 1. The difference between low and high thresholds in this comparison is only tens of millivolts. The netlists used were extracted from drawn layout and included all parasitic wiring capacitances and diode values. V_{DD} was set at 0.5 V and V_{SS} at 0.0 V. A temperature of 27 °C was used in the simulations. A series of four multiplexers was used in the simulations for each multiplexer type so that the outputs of each multiplexer would drive realistic loads. All unselected inputs were tied low.

Cascading an even number of multiplexers provided more realistic loads than just analyzing a single multiplexer, but it also helped to remove some of the state dependency in power consumption under DC conditions. While it is not a concern for high voltage processes (3.3 V for example), in low voltage design the subthreshold channel leakage dominates the DC current and increases with transistor width. If an inverter is sized so that it has similar pull-up and pull-down strength the P device will be larger than the N device due to the lower mobility of carriers in the P transistor. A logic ‘1’ at the input to an inverter turns on the N device while nearly turning off the P device. Under very low voltage operation, the P device will leak more because of its size than will a turned off N device. This causes measurably more DC current flow under a logic ‘1’ condition than under a logic ‘0’ condition.

3.1. BTS and non-BTS comparisons

Figures 5(a) and (b) show the schematics of the non-BTS and BTS multiplexer circuits. The areas of the two circuit layouts have been maintained, the transistor sizes are identical, and the input capacitances are constant. Because the BTS layout allows for an increase in some of the transistor sizes the transistors whose gates are controlled by signals Y1, Y1B, Y2, and Y2B were doubled in size to create another comparison circuit labeled “BigBTS”.

Table 1 shows rise and fall delays measured from $V_{DD}/2$ of the input signal to $V_{DD}/2$ of the output of the second multiplexer in the test series. The non-BTS circuit is significantly slower than the BTS circuits. This is primarily due to the added capacitance at node X. The BigBTS circuit was the fastest due to reduced impedance of the input signal path. The right side of the table was normalized to the fastest delay illustrating that the BTS circuit with near minimum

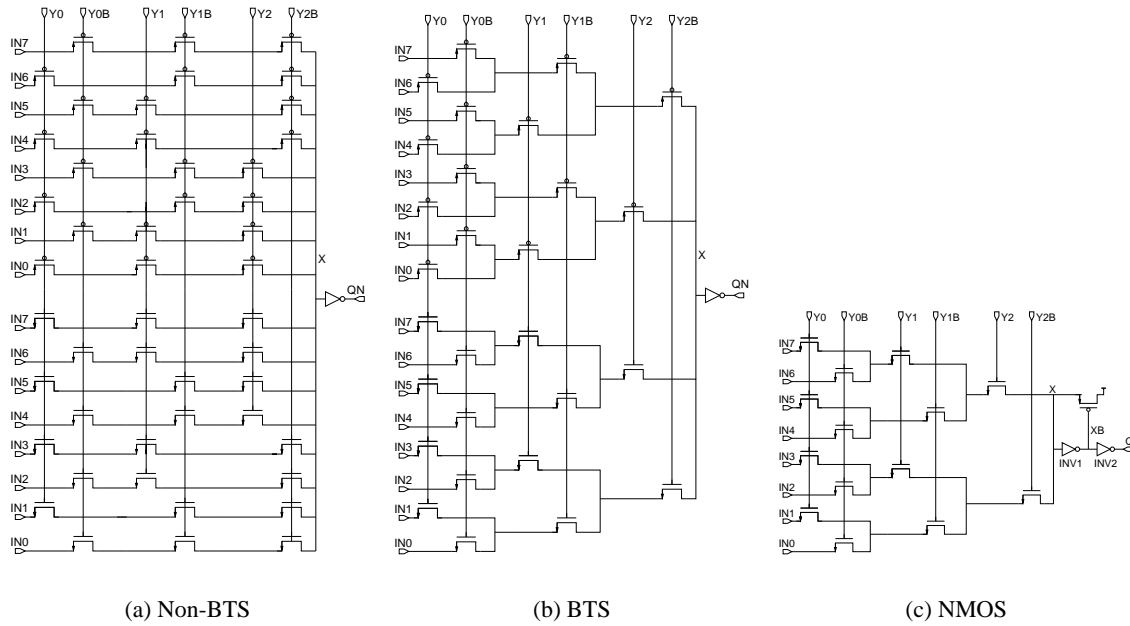


Figure 5. Multiplexers

Table 1. Rise/Fall delays. Right half is normalized

Vt	param	non-BTS	BTS	BigBTS	non-BTS	BTS	BigBTS
low	td1_qr	2.92 ns	1.56 ns	1.24 ns	2.35	1.26	1.0
low	td1_qf	2.45 ns	1.43 ns	1.26 ns	1.94	1.13	1.0
high	td1_qr	3.94 ns	2.16 ns	1.64 ns	2.40	1.32	1.0
high	td1_qf	3.45 ns	2.12 ns	1.65 ns	2.09	1.28	1.0

sized transistors was 13-32% slower than BigBTS while the non-BTS circuit was approximately 100% slower. As can be seen from the table, increasing the transistor threshold increases delay by approximately 40% for all multiplexer types.

Figure 6 shows the power versus frequency relationship for each of the multiplexer test cases at the low and high thresholds, as obtained from SPICE simulations of these cells. Power is for all four multiplexers. The BTS circuit consumes less power than the non-BTS. This is especially significant in the dynamic conditions in which the small input devices combined with high internal capacitance at node X increase dynamic current consumption. Even under static conditions, there is an increase in sub-threshold leakage of the non-BTS over the BTS circuit due to having more off transistors in the design.

What is interesting about the BigBTS multiplexer is that it consumes less dynamic power than the other multiplexers but consumes approximately 25% more static power than the others. The dynamic power reduction is due to faster rise and fall times at node X. The DC current increase may be explained in part because larger transistors exhibit greater subthreshold channel leakage current. The smaller BTS circuit exhibited the least leakage current.

It should be noted that in a typical 3.3 V process, using an identical netlist, the static power dissipation for the multiplexers is less than 2 nW. With high threshold values the sub-threshold channel leakage current isn't a major concern. With very low voltages, and hence very low thresholds, the subthreshold leakage currents dominate the static current results. With the low voltage

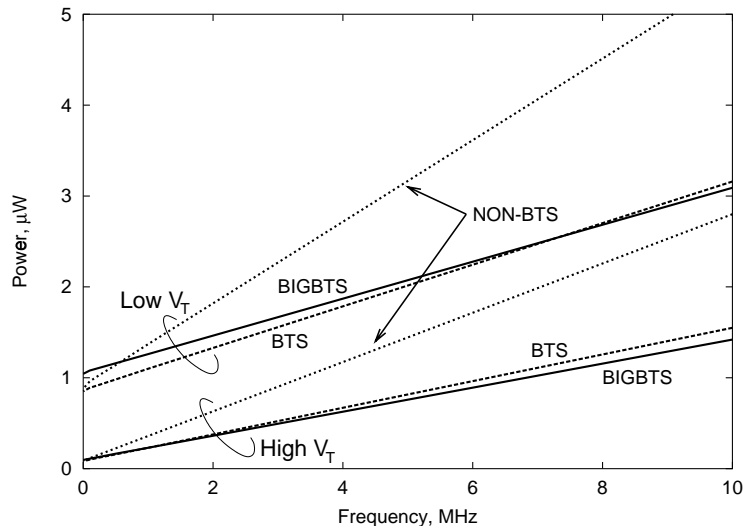


Figure 6. Power versus Frequency for Multiplexers with high and low threshold.

process, using a low threshold, the static currents are approximately 40 to 50 times greater than the static currents using the low voltage process with the higher threshold, and approximately 500 times greater than the static currents found in a typical 3.3 V process. The average dynamic power savings, however, is substantial when comparing a 3.3 V process with either of the low voltage process conditions: there is a reduction in dynamic power of a factor of approximately 35 by migrating to the low voltage process with the low threshold condition and a reduction of a factor of about 55 when migrating to the low voltage process with a high threshold.

3.2. NMOS multiplexer with a PMOS pullup

The NMOS multiplexer shown in Figure 5(c) has delay characteristics that are highly sensitive to threshold variations. The dynamic power consumption is very sensitive to pull-up strength. These two constraints make this circuit harder to design for low voltage applications than the multiplexers previously discussed that have full transmission gates.

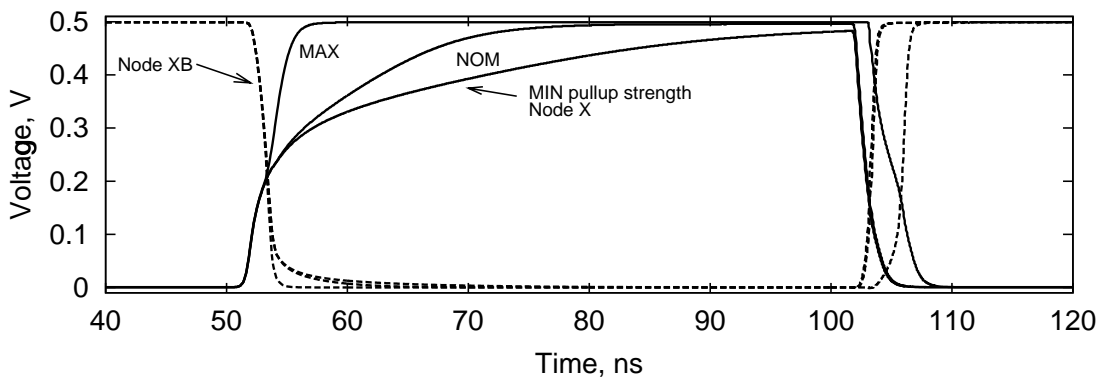


Figure 7. NMOS multiplexer node switching, Low threshold

Figure 7 and Figure 8 plot the voltage of internal node X as well as XB and Q over a cycle. Fig-

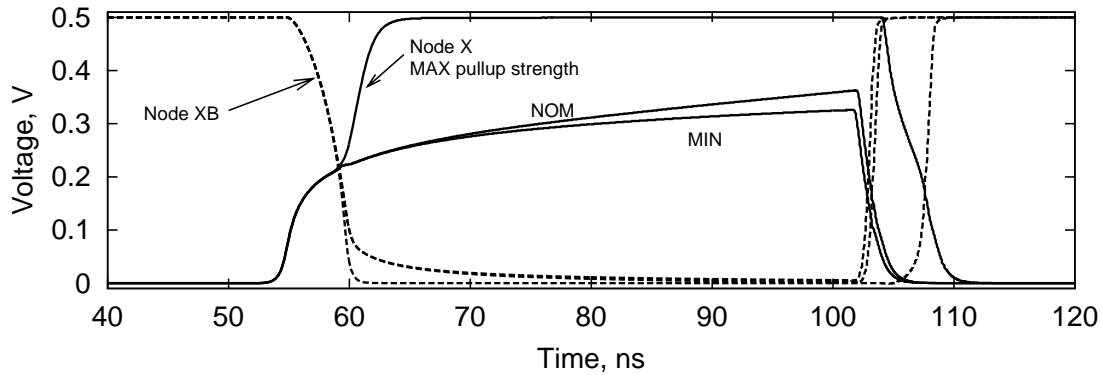


Figure 8. NMOS multiplexer node switching, High threshold

Figure 7 illustrates the low threshold condition and Figure 8 illustrates the high threshold condition. The simulation performed a sweep of channel length of the pull-up from minimum length, to 2 times, and 3 times minimum. INV1 was sized so its switchpoint was lowered to 200 mV to help compensate for the slow rise times through the NMOS devices. Under low threshold conditions, a minimum delay is achieved by having a weak pull-up. This is because the worst case delay is determined by the time it takes to pull-down the internal node X. A weak pull-up reduces the time for an input device to pull-down the internal node. This can be seen in Figure 7.

The high threshold case is shown in Figure 8. This plot shows the results of altering the threshold by approximately 50 mV. What is noticeable about the plot is that the fall time of node XB is much slower than the fall time of the low threshold case. The increased threshold slowed down charging node X. With the inverter threshold still at about 200 mV INV1 takes a long time to switch. To compensate for the increased transistor threshold, INV1's switching threshold could be lowered at the cost of noise margin. As in the case of the low threshold condition, having a stronger PMOS pull-up results in slower pull-down times at node X.

The difference in power consumption between the low and high threshold cases was only a few percent in the dynamic case (10 MHz) but nearly an order of magnitude under the static condition. The dynamic and static power consumption for the NMOS test cases was approximately 20 to 30% more than that of the multiplexers that used full transmission gates.

4. Low-voltage memory design

The design of a 10k x 16 single-access RAM is considered to illustrate the issues related to RAM design with low threshold transistors. The 10k x 16 Single Access RAM (SARAM) is a 10240 word x 16-bit static random access memory. The SARAM is split into 5 blocks of size 2k x 16. Each 2k x 16 block is divided into two blocks of size 1k x 16. Each 1k block is further divided into four sub-blocks of size 256 words. The 256 words are arranged as 32 rows x 8 columns. The SARAM is partitioned into smaller sub-blocks to reduce the leakage current effects on the bit-line as discussed in the next paragraph.

Figure 9 shows a 32-row stack consisting of 32 bit-cells (B0 to B31) and a write access controlled by WC with a precharge CLK. Assume the bit cell B0 is storing a logic '1' and all the other cells B1 through B31 are storing a logic '0'. When the address line A0 is active, the cell B0 is driving a logic '1' on the bit line D. All the other bit-cells (B1 to B31) are trying to pull the bit-line low due to the leakage current through the off transistors. If the number of rows is

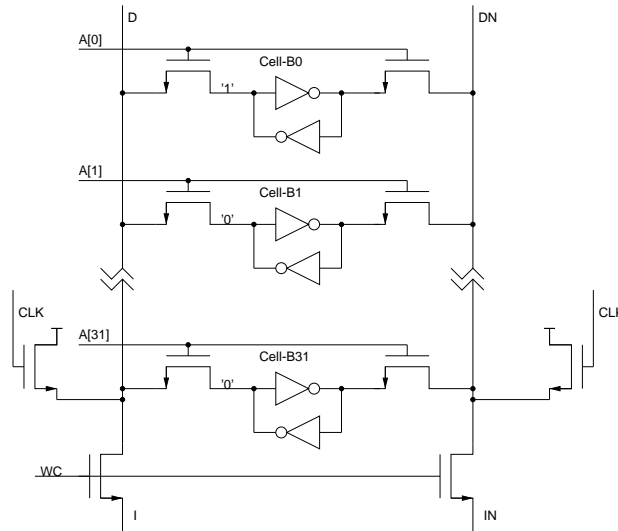


Figure 9. 32-Row Stack

increased then the increased leakage current can cause the bit-line to drift down to an intermediate voltage. Therefore, one of the first considerations in designing a low voltage RAM should be to determine the minimum number of rows that can be stacked and partition the RAM based on the minimum-row stacks. Reducing the number of the rows from the minimum-row stack will cause the address decoding and the read decoding logic to increase. This will slow the speed and also increase the power consumption. These tradeoffs must be carefully considered to determine the optimum partitioning of the RAM.

4.1. Comparison of the low threshold and the high threshold transistors

Another important factor in designing the RAM is the static power consumption of the non-switching nodes. We considered two design approaches, one using very low transistor threshold voltages and an operating voltage of 0.5 V and another with higher threshold voltages (over 100 mV) and a higher operating voltage of 1 V. For the high threshold case an operating voltage of 1 V was needed to provide a sufficient switching range for the bit-lines.

SPICE simulations were run to compare the power consumption of the SARAM for low threshold transistors at 0.5 V with the high threshold transistors at 1.0 V. Table 2 and the corresponding graph in Figure 10 show the comparison of the models for the power consumption in the 32-row stack in the Figure 9. Due to the symmetry and the sizing of the transistors, the power consumption results are not sensitive to whether a zero or one is stored in the bit cell.

Table 2. Power Consumption for 32-Row Stack

Frequency	Power	
	lower threshold, $V_{DD} = 0.5$ V	higher threshold, $V_{DD} = 1.0$ V
DC	5.03 μ W	1.65 μ W
20 MHz	5.18 μ W	3.53 μ W
30 MHz	5.38 μ W	4.61 μ W
40 MHz	5.60 μ W	5.60 μ W
50 MHz	5.81 μ W	6.61 μ W

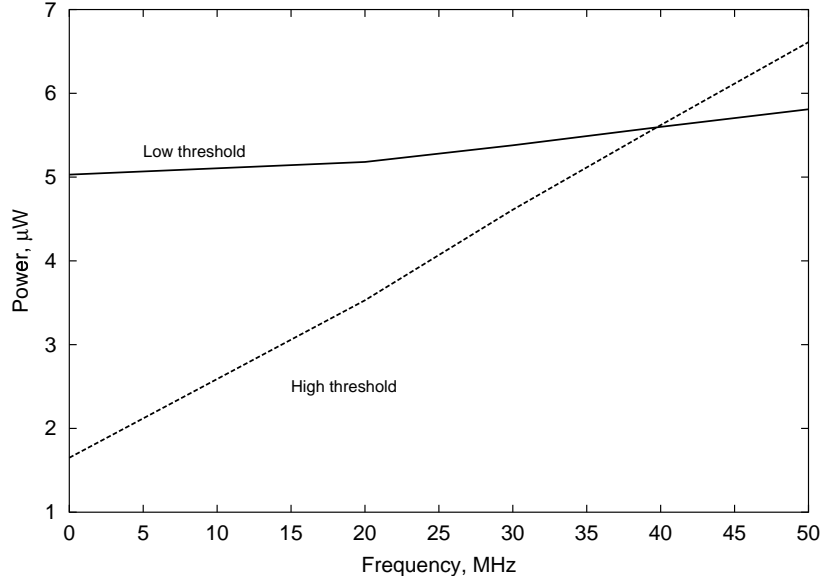


Figure 10. Power vs. Frequency for the 32-row Stack

The power consumption results of one 32-row stack operating at a frequency of 50 MHz are extended to all the bit-cells in the SARAM 10k x 16, and the resulting power consumption values are discussed in the next section.

4.1.1. Low threshold transistors

The smallest sub-block in the SARAM 10k x 16-bit is a 256 word x 16-bit block. The 256 words are arranged as 32 rows x 8 columns, so in any read cycle of the SARAM all the 8 columns of 16-bit each would be switching at the clock frequency. So 128 (8 x 16), 32-row stacks would be switching in one read cycle. The dynamic power consumption in the 128, 32-row stacks switching at 50 MHz frequency, based on the Table 2 is:

$$P_{DYNAMIC} = (5.81 - 5.03) \mu\text{W} \times 128 = 0.1 \text{ mW}$$

The total number of the 32-row stacks in the SARAM 10k x 16 is 5120 ($\frac{10240 \times 16}{32}$). Static power consumption in all the bit cells is:

$$P_{STATIC} = 5120 \times 5.03 \mu\text{W} = 25.73 \text{ mW}.$$

Total power consumption in the bit cells is:

$$P_{TOTAL} = P_{DYNAMIC} + P_{STATIC} = 0.1 \text{ mW} + 25.73 \text{ mW} = \underline{25.83 \text{ mW}}$$

4.1.2. High threshold transistors

Similarly, for the high threshold models, the dynamic power consumption in the 128, 32-row stacks switching at 50 MHz frequency, based on Table 2 is:

$$P_{DYNAMIC} = (6.61 \mu\text{W} - 1.65 \mu\text{W}) \times 128 = 0.63 \text{ mW}.$$

Static power consumption in all the bit cells is:

$$P_{STATIC} = 5120 \times 1.65 \mu\text{W} = 8.45 \text{ mW}.$$

Total power consumption in the bit cells is:

$$P_{TOTAL} = P_{STATIC} + P_{DYNAMIC} = 0.63 \text{ mW} + 8.45 \text{ mW} = \underline{9.08 \text{ mW}}$$

Due to very low activity factor in the RAMs the static power consumption is higher than the dynamic power consumption. The low threshold models with a 0.5 V operating voltage have a higher static current as compared to high threshold models at 1 V operating voltage. Based on the above discussion, it can be concluded that the RAMs should be designed with high threshold models with an operating voltage of 1 V. Level shifting interfaces can be used to interface the RAM to other logic operating at 0.5 V.

5. Conclusions

CMOS devices operating at very low supply voltages can provide dramatic reductions in power consumption. In order to maintain high performance levels it is necessary to reduce the transistor thresholds as well, and this can lead to slow switching, higher than expected power consumption, degraded logic levels, and even loss of functionality in some conventional circuit designs. Designers must consider the effects of increased subthreshold leakage, and should use circuits that are robust in the presence of threshold variations. When the activity level is quite low, as for a RAM, it may be more appropriate to choose a higher operating voltage, and higher thresholds, to minimize leakage currents.

References

- [1] Low-power electronics redraw satellite design. *Signal*, pages 91–94, August 2000.
- [2] Azeez J. Bhavnagarwala, Blanca L. Austin, Keith A. Bowman, and James D. Meindl. A minimum total power methodology for projecting limits on CMOS GSI. *IEEE Transactions on VLSI Systems*, 8(3):235–251, June 2000.
- [3] James B. Burr. Stanford Ultra Low Power CMOS. In *IEEE Low Power Workshop*, August 1993.
- [4] James B. Burr and Allen M. Peterson. Ultra low power CMOS technology. In *NASA VLSI Design Symposium*, pages 4.2.1–4.2.13, 1991.
- [5] Anantha Chandrakasan, Isabel Yang, Carlin Vieri, and Dimitri Antoniadis. Design considerations and tools for low-voltage digital system design. In *33rd Design Automation Conference*, June 1996.
- [6] Anantha P. Chandrakasan, Randy Allmon, Anthony Stratakos, and Robert W. Brodersen. Design of portable systems. In *Custom Integrated Circuits Conference*, pages 259–266, 1994.
- [7] Tsuneaki Fuse, Yukihito Oowaki, Takashi Yamada, Masahiro Kamoshida, Masako Ohta, Tomoaki Shino, Shigeru Kawanaka, Mamoru Terauchi, Takeshi Yoshida, Genso Matsubara, Shinichi Yoshioka, Shigeyoshi Watanabe, Makoto Yoshimi, Kazunori Ohuchi, and Sohei Manabe. A 0.5V 200MHz 1-stage 32b ALU using a body bias controlled SOI pass-gate logic. In *IEEE Solid State Circuits Conference*, pages 286–287, 1997.
- [8] Ricardo Gonzalez, Benjamin M. Gordon, and Mark A. Horowitz. Supply and threshold voltage scaling for low power CMOS. *IEEE Journal of Solid State Circuits*, 32(8):1210–1216, August 1997.
- [9] Dake Liu and Christer Svensson. Trading speed for low power by choice of supply and threshold voltages. *IEEE Journal of Solid State Circuits*, 28(1):10–17, January 1993.
- [10] Masayuki Miyazaki, Hiroyuki Mizuno, and Koichiro Ishibashi. A delay distribution squeezing scheme with speed-adaptive threshold-voltage CMOS (SA-Vt CMOS) for low voltage LSIs. In *International Symposium on Low Power Electronics and Design*, pages 48–53, August 1998.
- [11] Masayuki Miyazaki, Goichi Ono, Toshihiro Hattori, Kenji Shiozawa, Kunio Uchiyama, and Koichiro Ishibashi. A 1000-MIPS/W microprocessor using speed-adaptive threshold-voltage CMOS with forward bias. In *IEEE Solid State Circuits Conference*, pages 420–421, 2000.
- [12] Damu Radhakrishnan, Sterling R. Whitaker, and Gary K. Maki. Formal design procedures for pass transistor switching circuits. *IEEE Journal of Solid-State Circuits*, SC-20(2):531–536, April 1985.
- [13] Reto Zimmermann and Wolfgang Fichtner. Low-power logic styles: CMOS versus pass-transistor logic. *IEEE Journal of Solid-State Circuits*, 32(7):1079–1090, July 1997.