

Circuit Techniques for Gate and Sub-Threshold Leakage Minimization in Future CMOS Technologies

Rahul M. Rao, Jeffrey L. Burns*, Richard B. Brown
Dept. of EECS, University of Michigan, Ann Arbor, MI 48109

* IBM Austin Research Labs, Austin, TX 78758

<rmrao,brown@eecs.umich.edu,*jlburns@us.ibm.com>

ABSTRACT

Leakage current is becoming an increasingly important fraction of the total power dissipation of integrated circuits. This work focusses on leakage power minimization in light of the growing significance of gate leakage current. The need to consider gate leakage while determining the sleep-state pattern is demonstrated. Circuit reorganization and sleep-state assignment techniques are presented for gate and total leakage minimization of static and dynamic circuits. We also re-evaluate the MTCMOS circuit scheme for total leakage minimization.

1. INTRODUCTION

In recent years, the aggressive scaling of device dimensions and threshold voltage have significantly increased sub-threshold leakage and its contribution to the total chip power consumption. Also, gate oxide thickness has been scaled to maintain adequate control of the channel by the gate. This has resulted in an alarming increase of late in gate leakage current due to tunneling through the thin gate oxide. Gate leakage is expected to be a major component of leakage in future technology generations and has been identified as one of the most important challenges to future device scaling [1]. Gate leakage power, which was almost non-existent in the previous technology generations, is expected to contribute more than 15% to the total chip power dissipation in the 2004 technology generations.

To date, most circuit-level leakage minimization techniques focus only on sub-threshold leakage reduction, without considering the effects of gate leakage. Gate leakage is primarily being addressed from a CMOS technology perspective and the use of high-k gate dielectrics have been proposed. One of the approaches that addresses gate leakage, BGMOS [2], uses multiple threshold voltages and multiple oxide thickness devices. The use of PMOS dominated circuits was proposed in [3], on the basis that PMOS devices exhibit lower gate leakage compared to identical NMOS devices. However, due to band-to-band tunneling and use of different dielectrics [4], the gate leakage through PMOS devices is no

longer negligible and needs to be considered.

In this paper, we account for the contribution of gate leakage on total leakage by considering forward and reverse gate tunneling through both NMOS and PMOS devices. Gate leakage in conventional sleep-state patterns (which focus only on sub-threshold leakage) are evaluated and new sleep-state assignments for transistor stacks are proposed for total leakage minimization. We also present circuit re-organization schemes for total leakage reduction of dynamic circuits in sleep mode. Finally, we look into the effect of gate leakage on the MTCMOS circuit scheme and propose the use of sleep-state assignment in conjunction with MTCMOS to obtain increased total leakage savings.

The paper is organized as follows: An analysis of gate leakage and its dependencies is presented in Section 2. Section 3 describes sleep-state assignments for static CMOS circuits. Dynamic circuit re-organization schemes are presented in Section 4. Section 5 explores MTCMOS from a gate-leakage perspective. The findings and contributions are summarized in Section 6.

2. GATE LEAKAGE ANALYSIS

Figure 1 plots gate leakage current for an NMOS device as a function of gate-to-source (V_{GS}) and drain-to-source

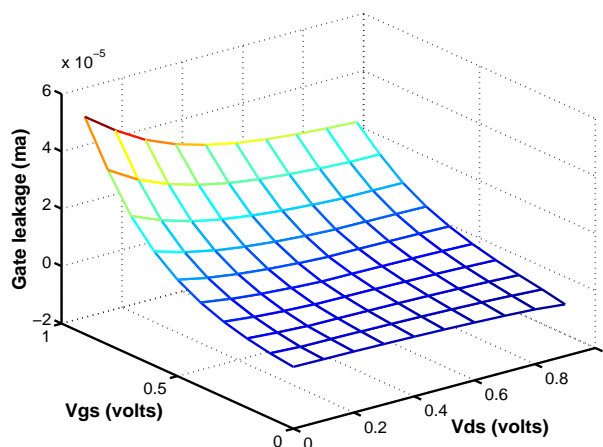


Fig. 1. Gate leakage current as a function of gate and drain bias for an NMOS device.

(V_{DS}) bias in a state-of-the-art sub 0.1 μm process. Gate leakage current shows an exponential dependence on the gate-to-source bias. At high gate bias, gate leakage current decreases with increasing drain-to-source bias. This can be attributed to the fact that a higher drain voltage results in a smaller electric field across the gate oxide at the drain end of the channel (lower V_{GD}). At low gate bias, gate leakage was found to increase with increasing drain bias (due to the increase in reverse gate leakage with increasing drain bias, i.e., V_{GD}). Thus, for a given gate-to-source bias, gate leakage is minimum when the gate-to-drain voltage is minimized. In addition, gate leakage current was found to be almost insensitive to the body-node voltage. The techniques presented in subsequent sections aim at minimizing the gate-to-source (V_{GS}) and gate-to-drain (V_{GD}) bias across a majority of devices, thereby obtaining a reduction in gate leakage and total leakage of the circuit.

3. STATIC CIRCUITS

Consider a three-high NMOS transistor stack (as found in the Nand3 cell shown in Fig. 2). The sub-threshold leakage through the transistor stack is minimized when all of the devices in the stack are turned ‘OFF’, i.e., when a $\langle 000 \rangle$ pattern is applied. Since conventional leakage-minimization techniques focus on sub-threshold leakage, the $\langle 000 \rangle$ pattern is believed to be the lowest leakage vector for a Nand3 cell. However, when such a pattern is applied, the output is high, and all of the PMOS devices experience high gate-to-drain and gate-to-source voltages. This results in a high field across the gate oxide causing gate leakage, which can be substantial due to the greater width of PMOS devices. To reduce gate leakage, it is necessary to maintain the terminals of most of the devices at the same potential. This can be achieved by turning ‘ON’ all but the lowest NMOS transistor in the stack, i.e., by applying the input pattern $\langle 110 \rangle$. Under such an input vector, only one PMOS device (P3) exhibits gate leakage. The gate leakage of the ‘ON’ NMOS transistors (N1, N2) is also negligible, since the internal nodes in the stack are charged almost to the supply rail (and hence the devices have a low V_{GS}/V_{GD}). The ‘OFF’ transistor (N3) at the bottom of the stack prevents sub-threshold leakage from increasing tremendously. Fig. 3 shows the sub-threshold, gate and total leakage for all possible input vectors for a Nand3 cell. The total leakage for some of these vectors is clearly dominated by the gate

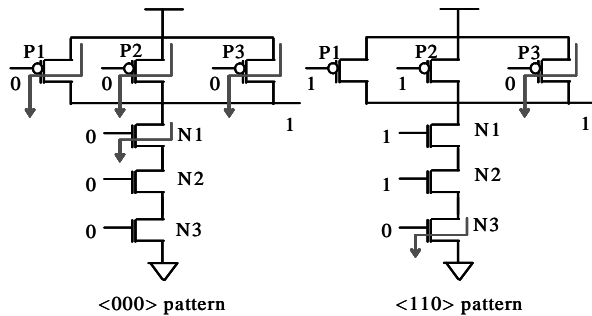


Fig. 2. Input patterns for total leakage analysis of Nand3 cell.

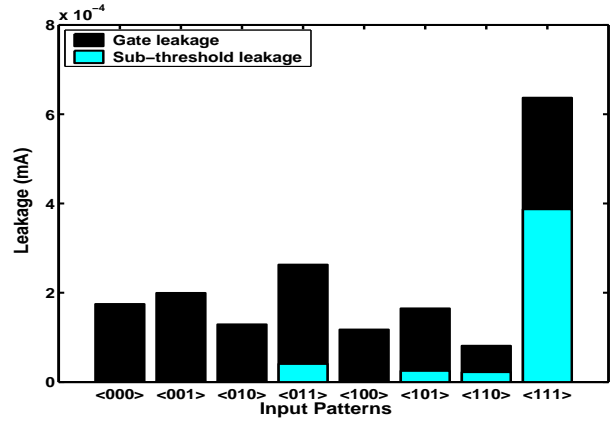


Fig. 3. Total leakage and its composition for all possible input vectors for a Nand3 Cell.

leakage component. Even though sub-threshold leakage for the vector $\langle 110 \rangle$ is greater than sub-threshold leakage for the vector $\langle 000 \rangle$, $\langle 110 \rangle$ is the minimum total leakage state for Nand3 cell. Thus, it is necessary to re-evaluate conventional leakage minimization schemes and input-vector assignments to account for the effect of gate leakage. With gate leakage expected to increase more rapidly than sub-threshold leakage, we expect that turning ‘ON’ all but the lowest device in a transistor stack will be the lowest leakage state for a transistor stack in future technology generations.

4. DYNAMIC CIRCUITS

This section focuses on sleep-state leakage minimization of dynamic circuits. Consider a typical 2-input dynamic And cell as shown in Fig. 4. During sleep state, the clock is held either in the precharge phase (low) or the evaluate phase (high). If the clock is held in the evaluate phase, the dynamic node will be discharged, and the output will be at logic high. Since, in a domino chain, the output of a dynamic cell drives other similar cells, it can be assumed that the inputs to the dynamic cell will also be at logic high. In such a state, all of the devices in the pull-down n-stack and the output pull-up transistor (i.e., devices on the evaluate path) will exhibit gate leakage. Since these devices are sized to reduce delay, it can result in significant gate leakage current. The sub-threshold leakage in this state is small, since it is primarily through the devices on the precharge path.

On the other hand, when the clock is held in the precharge phase, the dynamic node is charged high, the output will be at logic low and the inputs can be assumed to be at logic low. In this case, the devices on the precharge path exhibit gate leakage, while the devices on the evaluate path contribute to the sub-threshold leakage. Though sub-threshold leakage of the NMOS pull-down tree is minimal due to stacking effect, sub-threshold leakage through the wide output pull-up transistor can be considerable. Thus, in either of the two states, the total leakage of the cell may be high, although due to different mechanisms. Conventional techniques claim that holding the clock in the evaluate phase is the lowest leakage sleep state, but this approach completely neglects gate leakage.

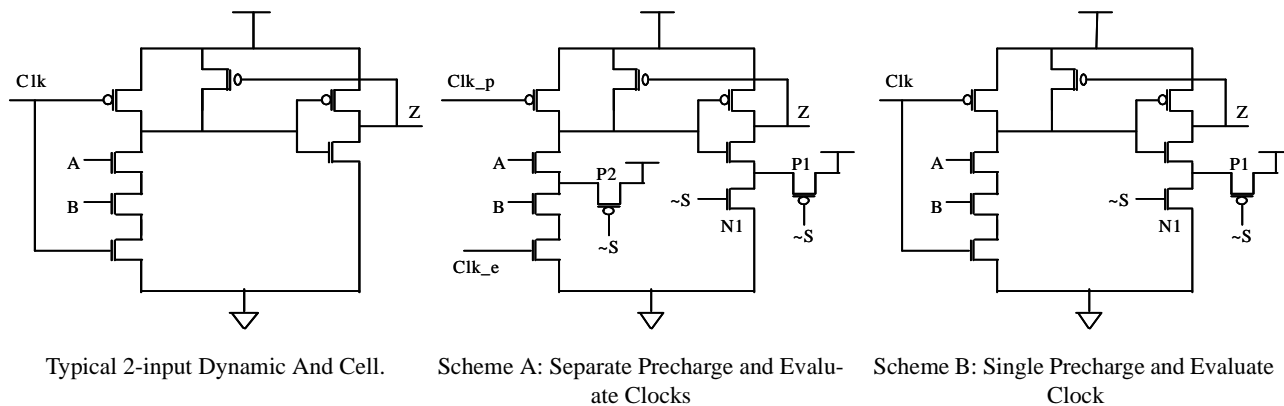


Fig. 4. Dynamic circuits: Circuit re-organization for gate and total leakage minimization.

Two proposed schemes are shown in Fig. 4. Both of these aim to minimize the total (sub-threshold plus gate) leakage current of the cell in sleep state. The output pull-down tree is modified to incorporate two small devices (N1, P1) that are controlled by the sleep-state control signal S. The precharge and evaluate clocks are separated in Scheme A. This will need additional circuitry for clock separation and can also result in clock skew problems. Scheme B uses a single clock, similar to the original configuration.

In Scheme A, in sleep state, the precharge clock is held high, while the evaluate clock is held low. The inputs to the cell can also be assumed to be high. The activated conditional pull-up devices (P1, P2) therefore charge the dynamic node and the output to logic high. In this state, gate leakage of both the evaluate and the precharge paths are reduced (only the evaluate transistor exhibits reverse gate leakage current) since most of the devices see an identical voltage at all of their terminals. The sub-threshold leakage of the output pull-up PMOS device is also reduced due to the ‘OFF’ device N1, resulting in significant savings in total leakage power. Since all of the additional devices are small, the delay degradation on the critical path is minimal. The additional devices can be desirably sized to obtain requisite precharge times and leakage savings.

For Scheme B, the clock is held low in sleep state. The dynamic node and the output of the cell are high (similar to scheme A) reducing the gate and sub-threshold leakage of the output PMOS inverter. The savings in total leakage is slightly reduced, since the precharge transistor exhibits gate leakage in addition to an increase in the sub-threshold leakage through the evaluate tree. However, in this configuration, no additional devices are needed in the evaluate tree, minimizing the delay degradation. The per-

centage savings obtained in gate leakage power and total power, along with the area overhead and degradation in precharge and evaluate times, are listed in Table 1 for several commonly-used dynamic circuits. For instance, savings of over 73% in gate leakage and 13% in total leakage are obtained for the dynamic And cell shown in Fig. 4 by using Scheme A with an area penalty of less than 7% and just over 1% degradation in delay.

5. MTCMOS CIRCUITS

The MTCMOS scheme has been proposed for reduction of sub-threshold leakage current in sleep state [5]. In this section, we investigate the effect of the MTCMOS configuration on gate and total leakage. The three configurations shown in Fig. 5 are considered. In the sleep state, the high V_T footer and header devices are turned ‘OFF’ (thereby minimizing sub-threshold leakage current). This causes the virtual supply rails to be close to V_{DD} or ground (if only footers or headers are used), or to be close to $V_{DD}/2$ (if both headers and footers are used).

The total leakage is the sum of the sub-threshold leakage of the sleep devices, the gate leakage of the sleep devices, and the gate leakage of the input stage. The devices of the first stage may exhibit gate leakage depending on the input vector. For instance, if only footers are used, the virtual ground plane will be close to V_{DD} . Thus, all of the devices in the logic circuit have their drain and source at nearly the supply rail. If an input vector of <0000.> is applied, then all of the devices in the first stage will see a high V_{GS} and V_{GD} , and hence exhibit gate leakage. However, when an input vector <1111.> is applied, these devices will have identical voltage at all of their terminals, resulting in minimal gate leakage. This makes the leakage in sleep-state for the

Table 1: Percentage savings and penalties for dynamic circuit re-organization schemes of Fig. 4 compared to conventional dynamic circuits.

Circuit	Scheme A					Scheme B				
	Gate leakage savings	Total leakage savings	Evaluate penalty	Precharge penalty	Area penalty	Gate leakage savings	Total leakage savings	Evaluate penalty	Precharge penalty	Area penalty
2 i/p And	73.67	13.71	1.27	28.18	6.08	51.6	14.3	-0.5	27.6	4.05
6 i/p 33Aoi	74.9	2.69	1.85	22.85	1.36	62.3	2.53	1.6	21.71	1.18
4-bit Mcarry	90.27	38.13	1.08	28.57	4.59	69.5	37.18	0.45	26.78	3.06

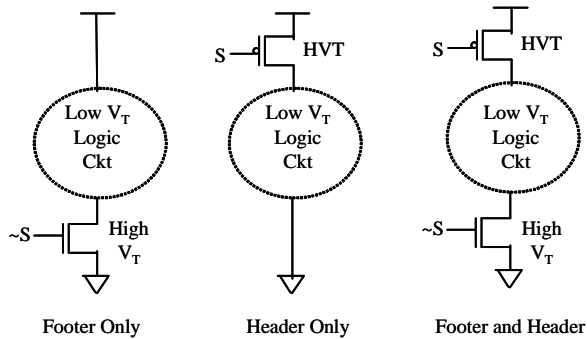


Fig. 5. MTCMOS configurations evaluated

footer-only configuration dependent upon the applied input vector. A similar argument can be presented for the header-only configuration. Hence, if a header-only or footer-only scheme is used, an appropriate input vector (00000... or 11111...) should be applied to obtain maximum savings in total leakage. This is validated in Fig. 6 which plots the gate leakage for an industry-standard decode circuit for each of the above schemes. In this case, the first 23 vectors are randomly generated, while vectors 24 and 25 are the <000...> and <111...> vectors. Here, the total leakage for the header_only configuration with the <000...> input vector is over 50% lower than the average leakage of remaining 24 vectors. Similarly, the application of the <111...> vector for the footer_only configuration results in over 40% savings compared to the average leakage for the other 24 vectors.

Further, the gate leakage of the sleep-state devices (headers/footers) can be significant, since these devices experience a high reverse V_{GS} and V_{GD} . Gate leakage can be reduced by using both headers and footers. When both headers and footers are used, the virtual supply and ground rails float close to $V_{DD}/2$. The gate-to-drain and gate-to-source bias across the sleep devices is reduced by about half, and hence their gate leakages are reduced. The gate leakage of the input stage is also reduced (and is less dependent on the input vector, as can be seen in Fig. 6).

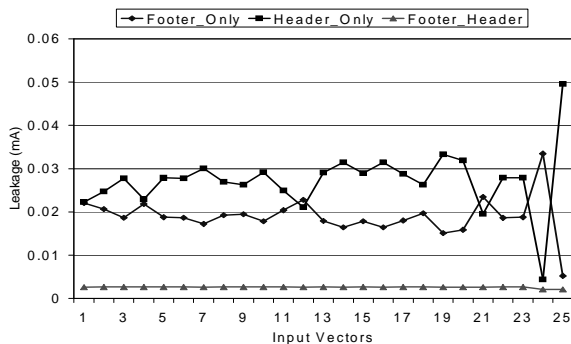


Fig. 6. Gate leakage as a function of input vector for MTCMOS circuit configurations of Fig. 5.

Table 2: Leakage currents ratio for MTCMOS configurations of Fig. 5 compared to the original circuit.

	Sub-threshold savings	Gate leakage savings	Total leakage savings
Footer Only	152.6	39.5	81.8
Header Only	111.8	28.9	61.6
Header and Footer	720.7	270.1	505.5

Table 2 lists the ratio of the leakage currents for the MTCMOS configurations of Fig. 5 compared to the leakage of the original circuit for an industry standard decode circuit in an advanced process.

6. CONCLUSIONS

In this paper we have shown the growing importance of gate leakage current and have clearly demonstrated the need to consider gate leakage in any leakage minimization scheme. We provided an analysis of gate leakage and presented optimal sleep-state assignments for transistor stacks in static circuits. We proposed new dynamic circuit configurations that result in over 70% reductions in gate leakage at a minimal performance penalty of less than 2%. The savings in total leakage current using these schemes range from 2% to 38% with less than 7% increase in device area. We also evaluated MTCMOS from a gate leakage perspective and illustrated the need to use both headers and footers to obtain maximum leakage savings. In cases where only headers or footers are used, we presented the optimal input vectors to be applied for reducing gate leakage which results in over 40% additional savings in total leakage current compared to random input patterns.

7. REFERENCES

- [1] International Technology Roadmap for Semiconductors, 2001 Edition, <http://public.itrs.net/Files/2001ITRS/Home.html>
- [2] T. Inukai, et.al, "Boosted Gate MOS (BGMOS): Device/Circuit Cooperation Scheme to Achieve Leakage-Free Giga-Scale Integration," *Proc. CICC 2000*, pp.409-412.
- [3] F.Hamzaoglu and M. Stan, "Circuit-Level Techniques to Control Gate Leakage for sub-100nm CMOS," *Proc. ISLPED*, pp. 60-63, Aug. 2002.
- [4] Y. Yeo, et.al, "Direct Tunneling Gate Leakage Current in Transistors with Ultrathin Silicon Nitride Gate Dielectric," *IEEE Electron Devices Letters*, vol.21, no.11, pp. 540-542, Nov. 2000.
- [5] S. Mutoh, et.al, "1-V Power Supply High-Speed Digital Circuit Technology with Multi-Threshold Voltage CMOS," *IEEE Journal of Solid State Circuits*, vol. 30, no. 8, pp. 847-854, Aug. 1995.