Sleep Transistor Design and Implementation – Simple Concepts Yet Challenges To Be Optimum

Kaijian Shi
Synopsys Professional Services
14911 Quorum Drive,
Dallas, Texas, USA
Kaijian.Shi@synopsys.com

David Howard
ARM Ltd.
110 Fulbourn Rd.
Cambridge, UK
David.Howard@arm.com

ABSTRACT

Optimum sleep transistor design and implementation are critical to a successful power-gating design. This paper describes a number of critical considerations for the sleep transistor design and implementation including header or footer switch selection, sleep transistor distribution choices and sleep transistor gate length, width and body bias optimization for area, leakage and efficiency.

INTRODUCTION

Leakage power has been increasing exponentially with the technology scaling [1][2]. In 90nm node, leakage power can be as much as 35% of chip power. Consequently, leakage power reduction becomes critical in low-power applications such as cell phone and handheld terminals. Power-gating is the most effective standby-leakage reduction method recently developed [3]-[6]. In the power gating, sleep transistors are used as switches to shut off power supplies to parts of a design in standby mode. Although the concept of the sleep transistor is simple, design of a correct and optimal sleep transistor is challenge because of many effects introduced by the sleep transistor on design performance, area, routability, overall power dissipation, and signal/power integrity. Currently, many of the effects have not been fully aware by designers. This could result in improper sleeper transistor design that would either fail to meet power reduction target when silicon is back or cause chip malfunction due to serious power integrity problems introduced. We have carried out comprehensive investigations on various effects of sleep transistor design and implementations on chip performance, power, area and reliability. In this paper, we shall describe a number of critical considerations in the sleep transistor design and implementation including header or footer switch selection, sleep transistor distribution choices and sleep transistor gate length, width and body bias optimization for area, leakage and efficiency.

A sleep transistor is referred to either a PMOS or NMOS high Vth transistor that connects permanent power supply to circuit power supply which is commonly called "virtual power supply". The sleep transistor is controlled by a power management unit to switch on and off power supply to the circuit. The PMOS sleep transistor is used to switch VDD supply and hence is named "header switch". The NMOS sleep transistor controls VSS supply and hence is called "footer switch". In sub-90nm designs, either header or footer switch is only used due to the constraint of sub-1V power supply voltage.

FINE-GRAIN VS. COARSE-GRAIN SLEEP TRANSISTOR IMPLEMENTATIONS

The sleep transistors can be implemented in a design in either "coarse-grain" or "fine-grain" power gating styles. In the "fine-grain" implementation, the sleep transistor is inserted in every standard cell which is often called MTCMOS cell. A power gating control signal is added to switch on and off power supply to the cell. An example of "fine-grain" NAND gate is shown in Fig. 1.

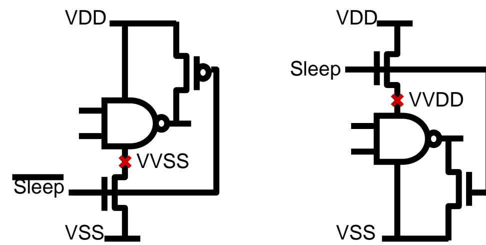

FIGURE 1. Footer and Header fine-grain sleep transistor implementation in NAND gate

A weak pull-up/down transistor controlled by the sleep signal is added to prevent floating output when the cell is in sleep mode. This is necessary to prevent short circuit current in those active cells connected to the sleep cell due to floating inputs. The pull-up/down transistor remains in OFF state in normal operation mode. Only one isolation state is allow which is "1" in footer switch implementations and "0" in the header switch implementations.

The advantage of the fine-grain sleep transistor implementations is that the virtual power nets (VVSS or VVDD) are short and hidden in the cell. Moreover, the MTCMOS cell can be implemented by existing standard cell based synthesis and place&route tools. However, the fine-grain sleep transistor implementation adds a sleep transistor to every MTCMOS cell that results in significant area increase. Also, it is not able to use the normal standard cells provided by library vendors and ASIC foundries. Another issue is that the MTCMOS cells become more sensitive to PVT variations, because the built-in sleep transistor is subject to PVT variation which results in added IR-drop variation in the cell and hence performance variation.

In the "coarse-grain" power gating designs as shown in Fig. 2, the sleep transistors are connected together between the permanent power supply and the virtual power supply networks.

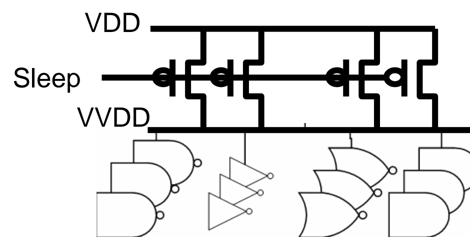

FIGURE 2. Header coarse-grain sleep transistor implementations

The main advantage of the "coarse-grain" power gating is that sleep transistors share charge/discharge current. Consequently, it is

less sensitive to PVT variation and introduces less IR-drop variations than the "fine-grain" implementations. Also, the area overhead is significantly smaller due to charge sharing among the sleep transistors.

Most power-gating designs prefer the "coarse-grain" sleep transistor implementation than the "fine-grain" implementation which incurs large area penalty and higher PVT sensitivity. In this paper, we shall focus on challenges in the "coarse-grain" sleep transistor designs and implementations.

HEADER VS. FOOTER SWITCH

The header switch is implemented by PMOS transistors to control Vdd supply. PMOS transistor is less leaky than NMOS transistor of a same size. The NBTI effect increases Vth over time and makes PMOS transistor even less leaky. Header switches turn off VDD and keep VSS on. As the result, it allows a simple design of a pull-down transistor to isolate power-off cells and clamp output signals in "0" state as shown in Fig.1. The "0" state isolation is complied with reset state requirement in most designs. The disadvantage of the header switch is that PMOS has lower drive current than NMOS of a same size, though difference is reduced by strained silicon technology. As a result, a header switch implementation usually consumes more area than a footer switch implementation.

The footer switch is implemented by NMOS transistor to control VSS supply. The advantage of footer switch is the high drive and hence smaller area. However, NMOS is leakier than PMOS and application designs become more sensitive to ground noise on the virtual ground (VVSS) coupled through the footer switch. The isolation on "0" state becomes complex due to loss of the virtual ground in sleep mode and necessity of bypassing footer switch to reach permanent VSS. In the following part of the paper, we shall focus on header switch design and implementations.

GRID VS. RING STYLE SLEEP TRANSISTOR IMPLEMENTATION

The sleep transistor has limited drive and relative high impedance compared with metal power rails. Consequently, sleep transistors are usually implemented as an array to provide sufficient drive current in a power gating design. The array can be implemented either in a ring style or a grid distribution.

In the ring style implementation, a virtual power ring is added to surround each power domain. The sleep transistors are placed between permanent power ring and virtual power rings to control power supply to each power domain, as shown in Fig, 3.

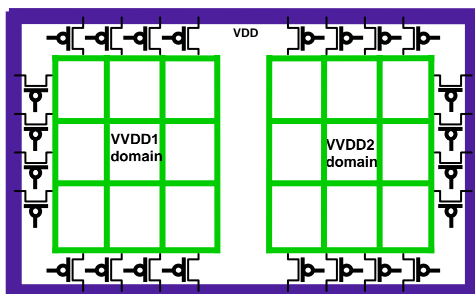

FIGURE 3. Ring style sleep transistor implementations

The ring style sleep transistor implementation is easy to implement and has small impact on placement and routing. However, it could result in more IR-drop at center of the design due to the limited drive of the sleep transistors distance from the center.

In the grid style sleep transistor implementation, the sleep transistors are placed close to power grid to connect permanent power network and virtual power networks, as shown in Fig. 4. The advantages of the grid style implementation are the better IR-drop management because each sleep transistor drives local cells. The sleep transistor distribution can be optimized to consume fewer sleep transistors than in the ring style implementation on a same IR-drop target. The drawback of the implementation is its impact on routing and physical synthesis, because the sleep transistors are distributed in the design area and their placement and routing constraints restrict layout optimization and net routing.

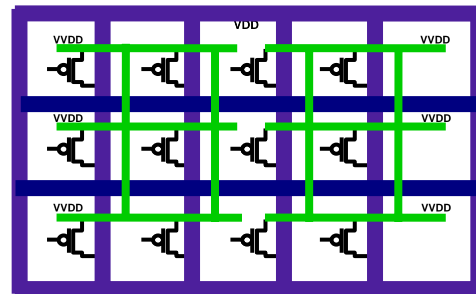

FIGURE 4. Grid style sleep transistor implementations

SLEEP TRANSISTOR DESIGN CONSIDERATIONS

The sleep transistor implementation introduces extra cost in chip area, routing resource, IR-drop and design complexity. There are also extra power dissipations from sleep transistors, power-gating control logic and power-on/off introduced operations. It is essential to ensure that the leakage reduction from the power gating implementation overwhelms those introduced costs to be worth the effort. To that end, various design considerations and tradeoffs need to be analyzed and handled correctly in the sleep transistor design and implementations. A good sleep transistor design is achieved by optimizing gate length and width, finger size and body-bias based on overall considerations of power efficiency, leakage current, IR-drop, area efficiency and layout impact.

SLEEP TRANSISTOR EFFICIENCY (Ion/Ioff)

The sleep transistor efficiency is defined by a ratio of drain current in ON and OFF states, i.e. Ion/Ioff. It is desirable to maximize the efficiency to achieve high drive in normal operation and low leakage in sleep mode. The sleep transistor efficiency can be analyzed by SPICE simulations where two high Vth transistors are configured for ON and OFF state respectively to measure Ion and Ioff. A high temperature is set on ON sleep transistor to model high chip temperature in operating mode and a low temperature is set on OFF sleep transistor to reflect the cool situation when the design is in sleep mode. The sleep transistor efficiency varies with gate length, width and body bias as shown by the curves in Fig. 5. The curves were generated by SPICE simulation of a TSMC90G high Vth PMOS transistor with foundry provided BSIM4 v2.0 model. The junction temperature of the transistor is set 125C^{o} in Ion analysis and 25C^{o} in Ioff analysis. Vds is set equal to Vdd in Ioff analysis and 10mV in Ion analysis reflecting the IR-drop target on the sleep transistor.

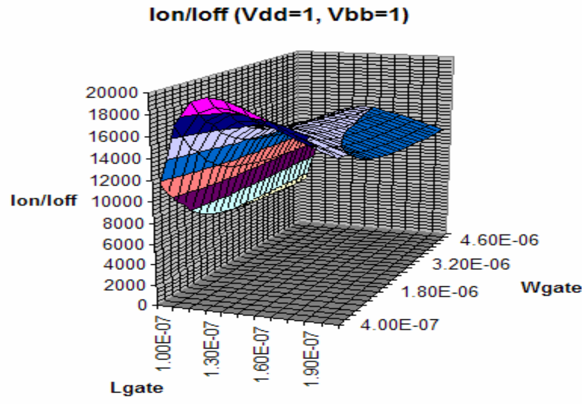

FIGURE 5. Ion/Ioff,-Lgate-Wgate curve

The sleep transistor efficiency increases with gate length (Lgate) and reaches peak at 130nm, mainly due to consequent Vth increase with Lgate and hence sub-threshold leakage current reduction. However, the efficiency declines after 130nm Lgate where Ion reduction with Lgate becomes more significant than leakage reduction. The efficiency also depends on gate width (Wgate). It drops quickly with increase of Wgate until Wgate reaches 1.6um. After that, it is level with Wgate. From efficiency point of view, a combination of long gate length at 130nm and small gate width is apparently a good choice.

The sleep transistor efficiency also depends on body bias because reversed body bias increases Vth and hence smaller sub-threshold leakage and higher efficiency. To evaluate the effect of body bias on the sleep transistor efficiency, we repeated the analysis above with various body biases. One of the results with 1.6V body bias is shown in Fig. 6.

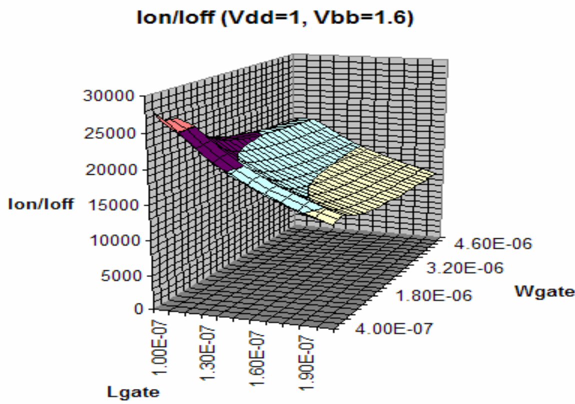

FIGURE 6. Ion/Ioff,-Lgate-Wgate curve with Vbb=1.6V

With 1.6V body bias, the sleep transistor efficiency increase by 40% compared with normal body bias where Nwell is connected to Vdd, i.e. Vbb=Vdd=1V. It is important to notice that the saddle shape Ion/Ioff curve in the normal body bias case disappears in the case of 1.6V body bias. The maximum efficiency occurs at close to process gate length which has higher drive current than the longer gate length (130nm) in the normal body bias case. Consequently, the sleep transistor of same drive current is smaller and more efficient with reversed body bias. However, further increase body bias beyond 1.6V will not improve the efficiency as shown by the solid line curves in Fig. 7 due to increase of body leakage and significant decrease of drain current.

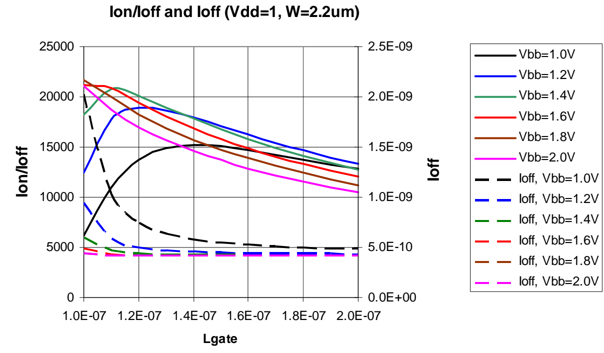

FIGURE 7. Ion/Ioff, and Ioff curves

Noticeably, the saddle point shift towards the process gate length with the increase of reversed body bias. This is because that reversed body bias increased Vth more effectively than by increasing gate length. At 1.6V body bias and above, Vth is mainly determined by the body bias and so is the subthreshold leakage current, as shown by the Ioff curves (dash lines) in Fig. 7. Although the reversed body bias requires extra power supply, it results in higher efficiency, stronger drive and smaller area sleep transistors. Therefore, it would be a better choice over the normal body bias in sleep transistor designs for ultra-low power applications.

IR-DROP CONSIDERATIONS

Besides Ion/Ioff efficiency, leakage current and drive current, IR-drop on sleep transistors must be considered in sleep transistor optimization in terms of gate length, width and body bias. IR-drop on the sleep transistor is tightly linked with equivalent channel resistance (Ron = Vds/Ids) when the sleep transistor is conducting. The smaller Ron, the smaller IR-drop. In a sub-50mV Vds region, Ron is linearly increased with gate length and body bias as shown by solid curves in Fig. 8. Ron is more sensitive to Lgate than Vbb. From the Ron and leakage curves in Fig. 8, we can see that at a same leakage current of 0.5nA (red and black dash lines), Ron is 1K Ohm (red solid line) in the sleep transistor of 100nm Lgate and 1.6V body bias compared with 1.5K Ohm (black solid line) in the sleep transistor of 180nm Lgate and normal (1V) body bias. It is clear the applying reversed body bias is a better choice than increasing gate length for Ron and leakage current reduction.

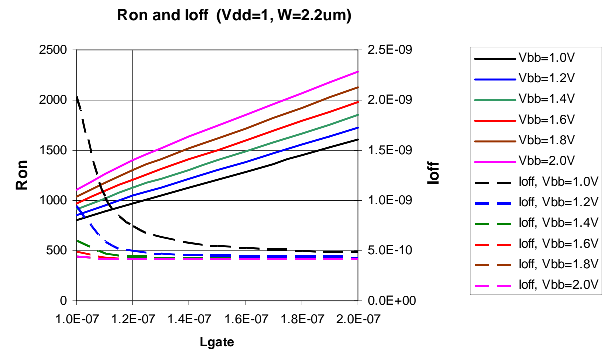

FIGURE 8. Ron and Ioff curves

Ron is also inversely proportional to Ids and Wgate as shown in Fig. 9. Consequently, Ron increases rapidly and hence is more sensitive to process variation when Wgate becomes smaller than 1.6um. This is contradiction to Ion/Ioff where smaller Wgate improve the efficiency.

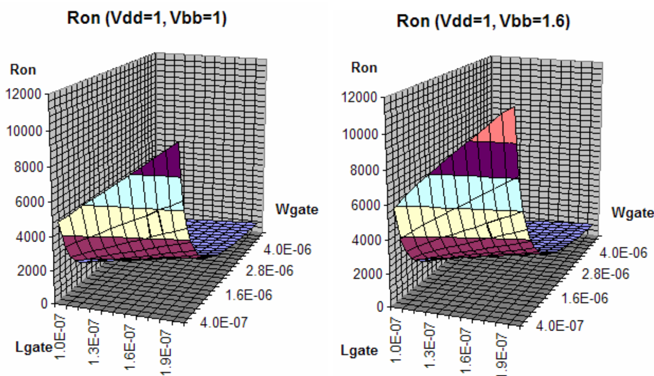

FIGURE 9. Ron curves with normal and reversed body bias

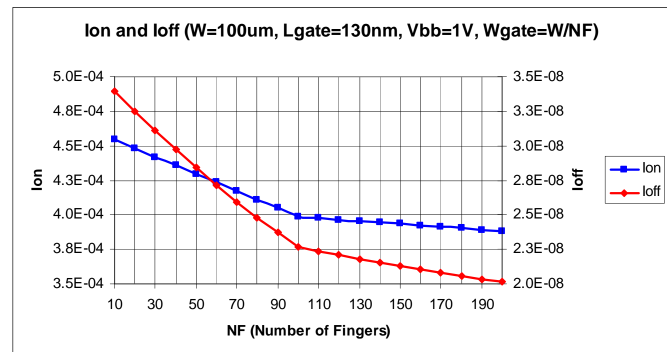

FIGURE 10. Ion and Ioff curves with multi-finger configurations

SWITCH CELL AREA EFFICIENCY

Area efficiency is another critical factor that must be considered in the sleep transistor design and implementation. The area penalty of the sleep transistors in a design can vary from 2% to 6% depending on how the sleep transistor is designed and implemented.

AREA EFFICIENCY IN SLEEP TRANSISTOR IMPLEMENTATION

Sleep transistors are implemented in an array as shown in Fig. 3 and Fig. 4. Given average current draw and IR-drop target of an application design, the total gate width of all the sleep transistors can be determined. The total gate width can be realized by various combinations of the number and gate width of the sleep transistors, i.e. fewer large sleep transistors placed in coarse grids or more small sleep transistors placed in fine grids. Considering the fact that minimum area overhead due to layout rule requirements occurs in a sleep transistor regardless the gate width, the fewer larger sleep transistor implementation is more area efficient than more smaller ones, because the minimum area overhead becomes less significant in a larger transistor. An example of such minimum area overhead is the minimum space between Nwell of the sleep transistor and other standard cells which have their Nwells connected to virtual Vdd rails. The hot Nwell spacing in 90nm node is about 0.6um. When a sleep transistor is abutted with a standard cell, the horizontal Nwell extension (0.3um) from the standard cells must also be considered. Consequently, 0.9um horizontal spacing is required at each side of the sleep transistor The vertical spacing is much smaller, because of built-in 0.3um vertical Nwell spacing in standard cell. Consequently, only 0.3um vertical spacing is required on the sleep transistor at each vertical side. The total spacing requirement on a sleep transistor is 1.8um in horizontal and 0.6um in vertical regardless the size of the sleep transistor. Although the hot Nwell spacing could be avoided by designing standard cells without connecting their Nwell to Vdd rails and connecting the Nwell to permanent Vdd at chip level in a 30um interval, other layout rules still impose minimum area overhead to the sleep transistor. Therefore, larger and fewer sleep transistors is more area efficient. However, the maximum size of the sleep transistor is constrained by impact on routability and IR-drop at center of a power grid. Once again, overall considerations are critical to an optimum sleep transistor design.

AREA EFFICIENCY IN SLEEP TRANSISTOR DESIGN

A sleep transistor is implemented in a multi-finger configuration in layout to provide sufficient current. For a given gate width of 100um, the Ion and Ioff vary with difference finger configurations as shown in Fig. 10.

The Ion drops with increase of number of fingers at 1.3% rate until 100 fingers of each 1um long. Then it reduces slowly at rate of 0.25%. Similarly, Ioff decreases fast at rate of 4.5% until 100 fingers and then slowly at rate of 1.4%. For high Ion/area efficiency, fewer longer fingers sleep transistor configuration is a good choice. However, if Ioff is also considered, the finger size of 1um is a better choice. Smaller than 1um fingers should not be considered because the area penalty in multi-finger transistor increases with number of fingers and the Ioff reduction is insignificant. The maximum finger size is limited by standard cell height and vertical spacing rules defined in a cell library. To improve area efficiency, the sleep transistor can be designed twice high as a normal standard cell in practice. This double high sleep transistor can be placed with standard cells with VDD and VSS rails aligned.

CONCLUSION

Although the concept of sleep transistor is simple, optimum sleep transistor design and implementation require optimizing all together the gate length, width and body bias with overall considerations of efficiency, leakage, drive, area and IR-drop effects which are often conflicting and need to be weighted based on application requirements. Increasing Lgate results in higher Vth and hence lower leakage and higher Ion/Ioff efficiency, at price of significant increase of Ron and decrease of Ion. Applying optimal reversed body bias is more efficient and effective alternative to produce a higher efficiency and Ion and lower Ron and Ioff sleep transistor than by increasing Lgate. Correct choices in sleep transistor implementations such as header or footer switch and ring or grid distributions are also important.

REFERENCES

[1] Kaushik Roy, Saibal Mukhopadhyay, and Hamid Mahmoodi-meimand, "Leakage current mechanism and leakage reduction techniques in deep-submicrometer CMOS circuits", Proc. IEEE Vol. 91, no. 2, Feb. 2003

[2] Dongwoo Lee, David Blaauw, and Dennis Sylvester, "Gate oxide leakage current analysis and reduction for VLSI circuits", - IEEE Trans. VLSI, Vol. 12, No. 2, Feb. 2004

[3] M. Powell, S.-H Yang, et. al. "Gated-Vdd: A circuit technique to reduce leakage in deep-submicron cache memories", in Proc. Int. Symp. Low Power Electronics Design, 2000, pp. 90-95

[4] Satoshi Shigematsu et. al., "A 1-V high-speed MTCMOS circuit scheme for power-down application circuits", IEEE J. Solid-State Circuits, vol. 32, no. 6, June, 1997

[5] Benton H Calhoun, Frank A Honore and Anantha P Chandrakasan, "A leakage reduction methodology for distributed MTCMOS", IEEE J. Solid-State Circuits, vol. 39, no. 5, May, 2004, pp. 818-826

[6] Changbo Long and Lei He, "Distributed sleep transistor network for power reduction", Proc. IEEE/ACM Design Automation Conference, 2003