Challenges in Sleep Transistor Design and Implementation in Low-Power Designs

Kaijian Shi
Synopsys Inc. (Design Services)
14911 Quorum Drive,
Dallas, TX 75254, USA
Kaijian.Shi@synopsys.com

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

ABSTRACT

Optimum power gating sleep transistor design and implementation are critical to a successful low-power design. This paper describes important 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. It also investigated various power-on current rush control methods for the sleep transistor implementation**.**

Categories and Subject Descriptors

B.0 [**Hardware General**]

General Terms

Design.

Keywords

low-power design, power gating, sleep transistor, methodology.

1. INTRODUCTION

Leakage power has been increasing exponentially with the technology scaling [1][2]. Power-gating is one of the most effective standby-leakage reduction method recently developed [3]-[7]. In a power gating design, sleep transistors are used as switches to shut off power supplies to parts of a design in standby mode. 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 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 and area penalty of the sleep transistors. Although the concept of the sleep transistor is straight forward, optimal sleep transistor design and implementation are challenge due to the needs of considering various effects, introduced by the sleep transistor and its implementations, on design performance, area, routability, overall power dissipation, and signal/power integrity. To make power gating worth the effort, the sleep transistors need to be optimally designed so that the benefit of leakage power reduction from power gating overwhelms the power and area penalties introduced by the sleep transistors and power-gating control circuit.

The implementation of the sleep transistor is also challenging. Adding more than necessary sleep transistors in a design will result in significant area penalty. Moreover, large power-on current rush when a design is coming out of sleep mode and charged by the sleep transistors simultaneously will cause large IR-drop in the design which in turn will cause malfunctions in the design due to IR-drop induced performance degradation and noise injections. The large current rush could also result in short term VDD collapse causing data corruptions in retention registers and memories.

We have carried out investigations on various effects of sleep transistor design and implementations on chip's performance, power, area and reliability. In the remainder of the paper, we shall describe critical considerations in the sleep transistor design including header or footer switch selection, sleep transistor distribution strategies, and sleep transistor gate size and body bias optimization for area, leakage and efficiency. We shall also analyze various power-on current rush situations, explain the causes and provide solutions to control the power-on current in the sleep transistor implementations.

2. CELL-BASED VS. DISTRIBUTED SLEEP TRANSISTOR IMPLEMENTATION

In the cell-based implementation, a sleep transistor is inserted in every standard cell which is often called MTCMOS cell. A power gating control signal is added to control the sleep transistor. An example of cell-based sleep transistor implementation of NAND gate is shown in Fig. 1.

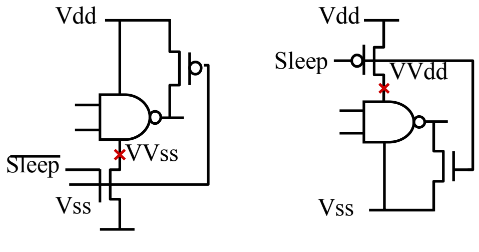

Figure1 1. Footer and Header cell-based sleep transistor implementation of an NAND gate

DAC 2006, July 24–28, 2006, San Francisco, California, USA.

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's output 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 cell-based sleep transistor implementation has two advantages. First, the virtual power nets (VVSS or VVDD) are short and hidden in the cell resulting in low sensitivity to noise injection and short power-on time. Secondly, the MTCMOS cell can be implemented by existing standard cell based synthesis and place&route tools. The disadvantages of the implementation are the large area penalty introduced by the sleep transistor in every MTCMOS cell and the cell sensitivity 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. Distribute sleep control signal to all the cells in a design is also challenging.

In the distributed sleep transistor implementation, the sleep transistors are connected between the permanent power supply and the virtual power supply networks. The main advantage of the distributed implementation is that sleep transistors can share charge/discharge current. Consequently, it is less sensitive to PVT variation and introduces less IR-drop variations than the cell-based implementations. Also, the area overhead is significantly smaller due to charge sharing among the sleep transistors. Most industrial power-gating designs adopt the distributed sleep transistor implementation. In the remainder of the paper, we shall focus on challenges in the distributed sleep transistor designs and implementations.

3. 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 consumptions in sleep transistors, power-gating control logic and power-on/off related operations such as saving and resorting states. To make power gating worth the effort, the sleep transistors need to be optimally designed so that the benefit of leakage power reduction from power gating overwhelms the power and area penalties introduced by the sleep transistors and power-gating control circuit. The sleep transistor is optimized in gate length, width, finger size and body-bias based on overall considerations of power efficiency, leakage current, IR-drop, area efficiency and layout impact.

3.1 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. 2. 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.

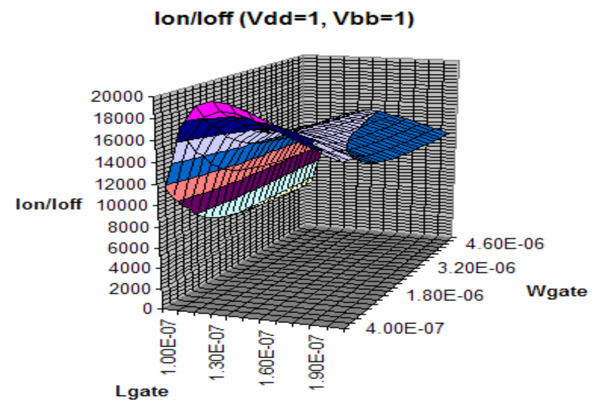

Figure 2. Ion/Ioff,-Lgate-Wgate curve with Vbb=Vdd=1V

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. 3.

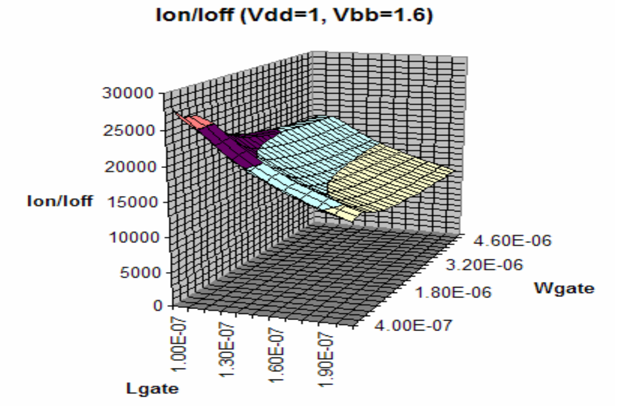

Figure 3. Ion/Ioff,-Lgate-Wgate curve with Vbb=1.6V

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. 4, due to increase of body leakage and significant decrease of drain current.

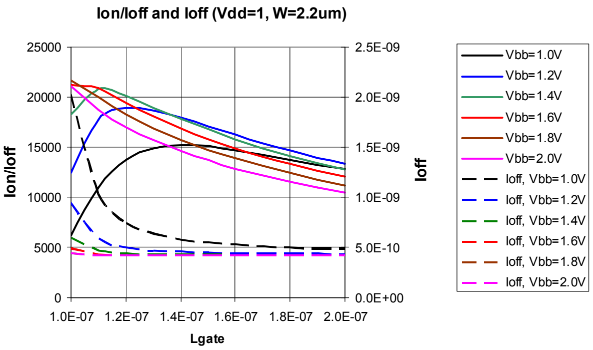

Figure 4. Ion/Ioff, and Ioff curves

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. 4. 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.

3.2 IR-drop considerations

Besides Ion/Ioff efficiency, leakage current and drive current, IR-drop on sleep transistors must be considered in sleep transistor optimization. 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. Ron is more sensitive to Lgate than Vbb. Applying body bias results in 30% smaller Ron than by increasing gate length with a same leakage current target.

3.3 Area efficiency

Area efficiency is another critical factor in 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. 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 sleep transistors and gate width of each sleep transistor. 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 placed in coarse grids is more area efficient than more smaller sleep transistors placed in fine grids, because the minimum area overhead becomes less significant in a larger transistor. 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.

A sleep transistor is implemented in a multi-finger configuration in layout to provide sufficient current. The choice of number of fingers for a given sleep transistor gate width also affects the area efficient. For high Ion/area efficiency, a sleep transistor configuration with fewer longer fingers is a good choice. However, if Ioff is also considered, the smaller finger size would result in higher Ion/Ioff. 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 as high as a standard cell.

4. POWER-ON CURRENT RUSH CONTROL METHODS

4.1 Chained power-up method

A solution to control power-on current is to turn on the sleep transistors consecutively in a daisy chain style. In this case, the power-on current gradually increases with the number of turn-on sleep transistors. However, the current rush resulting from this method could still be too large unless the daisy-chain is very slow so that the power rail has enough time to be slowly charged close to VDD when all sleeper transistors are turned on. The delay through a typical chain of inverters, for example, is likely to be much shorter than the time required to power up a large voltage island, so it is likely that the peak value of the current will be large. On the other hand, if on-chip resistive elements, such as a long polysilicon resistor, to produce a larger propagation delay in the chain, then the slowly rising voltages on circuit inputs could introduce other problems such as crow-bar currents and hot electron effects.

4.2 Two stage power-on method

Another approach is to split the sleep transistors into two arrays: a weak transistor array and main transistor array. At power on, the weak transistors are turned on first to trickle charge the design. The limited current-flow of the weak sleep transistors constrains the power-on current rush. When the design is charged to a voltage close to VDD, the main header transistor array is turned on ready for normal operation. One possible control mechanism for the turn-on sequence of both arrays is a daisy chain of buffers, as shown in the diagram in Fig. 5, which once again spreads out the turn-on time and reduces dI/dt.

The size of the weak trickle transistors is determined by the defined maximum current limit of the design although it may also need to consider the maximum permissible turn-on delay time. The smaller transistors have lower current peak, but take longer for the switched power supply to reach 90% of the final operating voltage. The designer must decide how much current flow is acceptable, based really on how much voltage perturbation is acceptable at the storage elements. Another consideration is how those weak header transistors are distributed across the cell - they may be cluster together in a single region or column for easy control, or they may be scattered across the whole layout to minimize local IR drop.

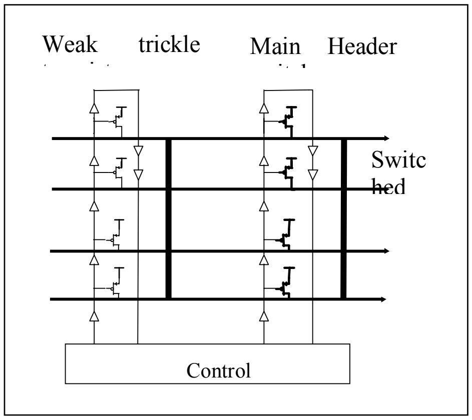

Figure 5. Two stage power-on daisy chains.

4.3 Main header turn-on control

Once the weak transistor array configuration has been determined, we need to decide when to turn on the Main Header array, based on the design constraints on peak current and power-on latency.

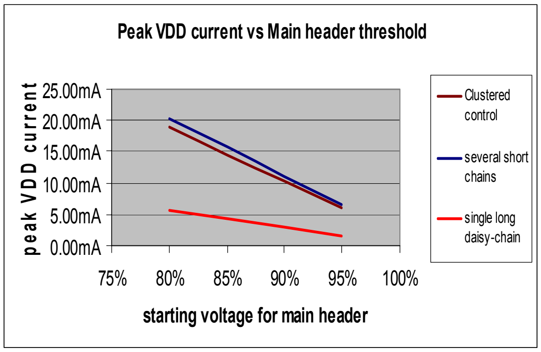

Figure 6. Peak power-on current with difference threshold

The curves in Fig. 6 show the peak VDD current for the test-case with three different Main Header configurations at different main sleep transistor turn-on threshold. In the first example, the Header transistors are bundled into a few of clusters which are distributed across the test macrocell. Sleep transistors in a cluster are turned on at the same time. The clusters are connected in a chain and tuned on in sequence. This configuration gives rise to the highest peak current. In the second example, the Header transistors are connected in several short daisy-chains spread out across the macrocell and turned on in quick succession. This also leads to a fairly high peak current. In the third configuration, all the Headers are linked in a long daisy-chain running all around the macrocell. This configuration leads to the lowest peak current, but takes much longer to charge VDD to the operating voltage. In all cases, the peak current decrease linearly with the increase of the turn-on threshold. If power-on latency budget were not tight, the single chain main header configuration with turn-on threshold of 95% of VDD would prevent power-on current rush. On the other hand, two to four daisy chains would be a compromise to meet both peak current and power-on latency budgets. The main header turn-on can be controlled by a simple delay circuit using the weak transistor daisy-chain to model the time to trickle charge the rail to the turn-on threshold. Two drawbacks of the method are the need to analyze and create required delay for each design and the significant delay variation in the long chain. An alternative scheme is to employ a voltage detector such as a Schmitt trigger which switches at the desired turn-on threshold voltage regardless the size of the sleep transistor array.

5. CONLUSION

Although the concept of sleep transistor is simple, optimum sleep transistor design and implementation are challenge. They require optimizing gate length, width, body bias and finger configuration with overall considerations of efficiency, leakage, drive, area and IR-drop effects which are often conflicting and need to be weighed based on application requirements. Increasing Lgate results in higher Vth and hence lower leakage and higher Ion/Ioff efficiency, at the 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.

Current rush at power-on is a critical issue in the sleep transistor implementation. It can cause large IR-drop and short term VDD collapse resulting in malfunctions in the design. Among various current rush control methods, the two stage charge method is most effective. The size and number of the weak transistors are largely determined by the power-on current rush limit. The main header turn-on can be controlled by a design dependent delay circuit using the weak transistor daisy-chain or a voltage detector such as a Schmitt trigger. The main headers can be configured as clusters or daisy chains which are turned on in sequence. The configurations and optimal number of chains are determined based on max peak current and power-on latency budget.

6. 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

[7] Anand Ramalingam, Bin Zhang, Anirudh Davgan and David Pan, “Sleep Transistor Sizing Using Timing Criticality and Temporal Currents”, Proc. ASP-DAC, 2005