Low Power Integrated Scan-Retention Mechanism

Victor Zyuban
IBM T.J. Watson Research Center,
Yorktown Heights, NY
zyuban@us.ibm.com

Stephen V. Kosonocky
IBM T.J. Watson Research Center,
Yorktown Heights, NY
stevekos@us.ibm.com

ABSTRACT

This paper presents a methodology for unifying the scan mechanism and data retention in latches which leads to scannable latches with the data retention capability achieved at a very low power overhead during the active mode. A detailed analysis of power and area overhead is presented, with layout examples for various common latch styles. Implications of using different power gating techniques for reducing leakage during sleep mode on the design of retention latches are considered, including well biasing for leakage control and sharing wells between gated logic and retention latch devices.

Categories and Subject Descriptors

B.2.1 [**Design Styles**]: Pipeline; B.6.1 [**Design Styles**]: Sequential circuits; B.7.1 [**Types and Design Styles**]: VLSI

General Terms

Design

Keywords

data retention, MTCMOS, subthreshold, leakage, low power, latch, scan, balloon latch

Introduction

As CMOS process technology is scaling, power supply voltage scales down as well, and so do transistor threshold voltages to maintain high-speed operation. Although lowering the threshold voltage reduces circuit delays, it also exponentially increases the subthreshold leakage currents. These leakage currents lead to power dissipation even when the circuit is not doing any useful computations which presents a serious problem for battery operated devices.

The complexity of modern designs has reached a point where saving power by implementing a non-scannable design is not viable. Scannable latches are part of the standard testing methodology, however traditional methods for implementing scannable designs have a power overhead that is not negligible [11]. In this paper we propose an integrated scan-retention mechanism that has a much smaller total power overhead in the active mode than the combination of the prior art scan and retention mechanisms, implemented independently. We show how the proposed mechanism can be applied to a variety of latch styles, including those recently reported. We prove the concept with a developed layout and test chips manufactured in a state of the art 0.13μ technology, and measure the exact values for the power and area overheads of the proposed mechanism.

ISLPED'02, August 12-14, 2002, Monterey, California, USA.

1. PRIOR ART BALLOON LATCH RETENTION MECHANISM

A standard method of reducing the leakage power during inactive intervals is to use the multi-threshold CMOS (MTCMOS) technology, together with sleep or power down modes. According to this methodology, all logic is built of low-threshold transistors, with a high-threshold transistor serving as a footer or a header to cut leakage during the quiescence intervals. During the normal operation mode, the MTCMOS circuits achieve high performance, resulting from the use of low-threshold transistors. During the sleep mode, high threshold footer or header transistors are used to cut off leakage paths, reducing the leakage currents by orders of magnitude. During the power-down mode the state of all circuits, connected to the power supply (or ground) through the header or footer is lost. In most cases the state of the circuit needs to be restored on returning from the power-down mode, to resume the operation. The state of sequential circuits is stored in latches or flip-flops, consequently, to resume the operation of the sequential circuit after returning from the standby mode, the state of all latches or flip-flops needs to be restored. A prior art technique for saving and restoring the state of latches during the power-down mode in MTCMOS sequential circuits is based on duplicating every regular latch or flip-flop in the circuit with a shadow or balloon latch, and providing a path to move data from the regular flip-flop to the shadow, and back [2, 7]. The balloon, or shadow latch, shown in Fig. 1, is built of high-threshold devices, and connected to real power and ground (bypassing the footer and header transistors). Since the leakage currents through the high threshold devices are orders of magnitude smaller than those through the low-threshold transistors, the leakage currents through the balloon latch during the power-down mode are small, and can be neglected.

Fig. 1a shows that adding the balloon latch adds ten extra transistors, to the flip-flop, increasing the transistor count from 16 to 26. Two inverters and transmission gate T7, comprising the balloon latch add 6 transistors to the circuit. Transmission gates T5 and T6 add 4 more transistors to the circuit, that provide the path for moving data between the main latch and the balloon latch. Thus, the transistor count overhead of the balloon latch is at estimated

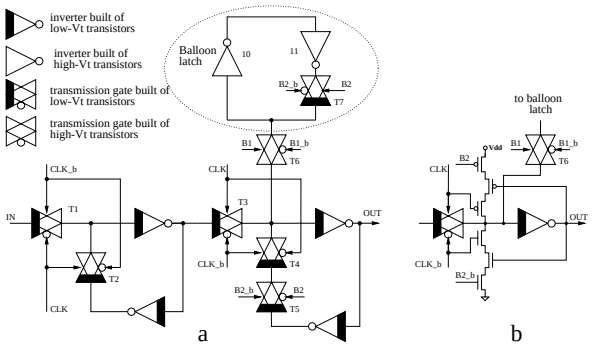

Figure 1: Prior art balloon latch.

as $10/16 = 63\%$ (the actual area overhead may be smaller though, because of the small size of transistors in the balloon latch).

Adding the balloon latch also leads to an increase in delay through the latch and its active power, because of the extra parasitic capacitance of the two transistors of transmission gate T6 that gates data to and from the balloon latch, and the two transistors in transmission gate T5 in the feedback path of the slave latch. For the minimum size transistors, the total introduced capacitance that is charged/discharged during the active mode of operation is $C_b = 6C_d + 2C_g + C_w$, where C_d and C_g are source and drain capacitances of minimum size transistors, and C_w is the wiring capacitance overhead. This introduced capacitance switches every time new data is latched. For a latch working with non-overlapping phases of the clock, it is not affected by glitches at the latch input. Thus, the power overhead of the data retention feature in the active mode can be estimated as $P_r = \frac{1}{2}\alpha f C_b V_{dd}^2$, where f is the clocking rate and α is the activity factor. Simulations show that for a 0.13μ bulk technology with $V_{dd} = 1V$, $f = 250$MHz and $\alpha = 0.3$ the power overhead of the balloon latch in a design that has 4000 latches is approximately $P_r = 0.5mW$, which is not negligible in state of the art low power designs. If the feedback path in the slave latch is built as shown in Fig. 1a, however, the tranistors in the transmission gate T5 do not introduce any capacitance that switches during the normal operation mode, which reduced the introduced capacitance C_b to $C_b = 2C_d + C_w$

If there is the scan requirement in addition to the data retention, and the scan mechanism is implemented independently of the retention mechanism, then the combined overhead of the scan-retention mechanism gets more significant. Even for the state of the art low-power scan mechanism in [13] the power overhead of the scan feature, for the same assumptions is $P_s = 0.3mW$, and the total power overhead of the retention and scan mechanisms, implemented independently is $P_{s+r} = P_s + P_r = 0.8mW$, which may exceed 5% of the active power of state of the art low power low-power product of this size.

In this paper we propose an integrated scan-retention mechanism that has a much smaller total power overhead in the active mode than the combination of the prior art scan and retention mechanisms, implemented independently. We show how the proposed mechanism can be applied to a variety of latch styles, including those in the recently reported works. We prove the concept with a developed layout and test chips manufactured in a state of the art 0.13μ technology, and measure the exact values for the power and area overheads of the proposed mechanism.

2. NEW INTEGRATED SCAN-RETENTION MECHANISM

The proposed integrated scan-retention mechanism is based on a prior art low-power overhead level-sensitive scan mechanism [13], shown in Fig. 2. The master latch in Fig. 2 can be any type of a single phase latch, or a two-phase latch, for example, edge-triggered latch, pulsed latch [8, 10], or dual edge triggered flip-flop [5, 3].

The scan latch is a low-area slow level-sensitive latch, controlled by clock B. The output of the scan latch is the scan output of the entire flip-flop. It is connected to the scan input of another latch in the scan chain. During normal operation mode, clock A and clock B are kept at the low level, and the flip-flop works as a conventional latch, whereas scan latch is in the non-transparent state, so that the scan output does not toggle, and the internal capacitances inside the scan latch do not toggle either. This reduces the power dissipation in the normal operation mode. During the scan mode, clock C is kept at the low level, and the flip-flop works as a master-slave latch, controlled by non-overlapping clocks A and B, providing a robust, level-sensitive scan operation.

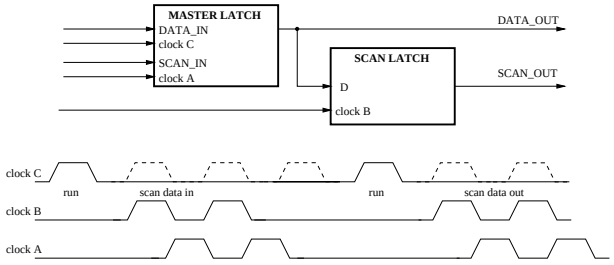

Figure 2: Prior art low power scan mechanism.

Fig. 3 and 6b show implementation examples of the above scan mechanism, applied to some of the recently published latches [8, 10, 5, 3]. For abutted latches in custom datapaths scan outputs **SO** and **SO_b** of a latch are connected to scan inputs **SI** and **SIb** of an adjacent latch. For distant scan connections a single-rail connection is used, and input **SI** is locally inverted. Such a connection scheme reduces the transistor area overhead of the scan mechanism without incurring significant routing overhead.

The power overhead of this scan mechanism is reduced to the gate and drain capacitance of two minimum-sized transistors, connected to the output nodes plus some wiring overhead. This extra capacitance is charged or discharged at most once per clock cycle, and is not affected by spurious transitions at the data input.

Fig. 4 shows the proposed extension to the above scan mechanism to provide the low-overhead data retention capability. The new flip-flop with retention uses the scan latch as a low-leakage storage for retaining data during the power-down mode, and provides an extra path for restoring the data from the retention latch to the main flip-flop. The retention latch is built of low-leakage devices, such as high threshold transistors, or regular transistors with the back bias capability. If gate leakage is an issue, transistors in the retention latch can be implemented as thick gate oxide devices. Real ground and Vdd are used as power terminals in the retention latch. The rest of the flip flop is built of fast low threshold, thin gate oxide transistors, and it may use either virtual Vdd with a header, or virtual ground with a footer, to cut the leakage path during the power-down mode.

During the normal operation mode clocks A and B are kept at the low level, and the latch operates as the conventional latch. During

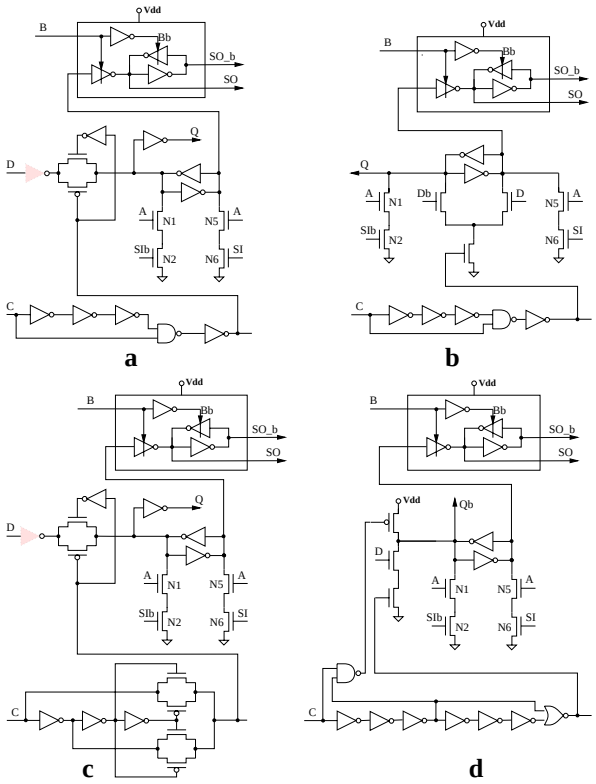

Figure 3: Scannable latches: a - ep-SFF; b - SSASPL; c - ep-DSFF; d - DPSCRFF.

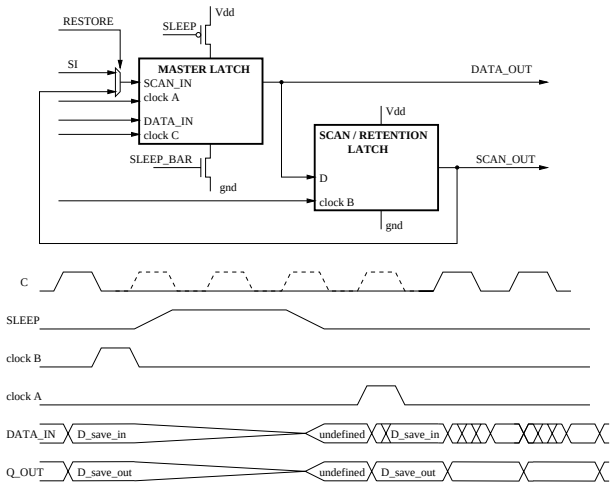

Figure 4: Scannable latch with data retention.

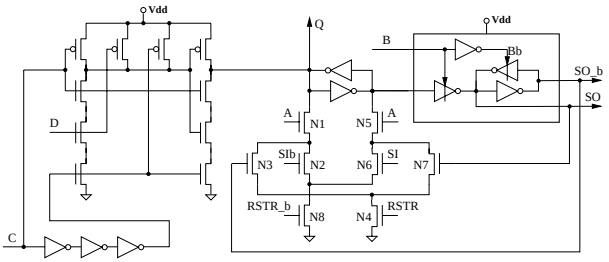

Figure 5: Scannable HLFF latch with data retention.

the scan mode the RESTORE signal is kept at the low level, and the latch work as a master-slave latch, controlled by clocks A and B, as described earlier. When entering the sleep mode, high level at clock B saves data in the retention latch, Fig. 4. On returning from the sleep mode, high level is applied to the RESTORE signal, and high level at clock A restores data from the retention latch to the main flip-flop.

Fig. 5 and 6d show examples of implementing the proposed data retention mechanism in the scannable HLFF and sense amplifier latches. The path for restoring data from the retention latch to the main latch is implemented as a stack of NFETs N3 or N7 and N4. The proposed scan-retention mechanism can be applied to a variety of latches [1, 4, 8, 9], including those in Fig. 3.

The additional power and delay overhead of the retention mechanism over the scan mechanism is reduced to a minor increase in capacitances of internal wires, due to some increase in the area of the flip flop. No extra capacitance of transistor gates, sources or drains is added to any nodes that are switching during the normal operation mode. This feature makes the proposed retention mechanism particularly attractive for low-power applications, where minimizing both active and standby power is important

3. ANALYSIS OF THE OVERHEAD OF DATA RETENTION

To determine the area, delay and power overheads of the proposed data retention mechanism we developed the layout in 0.13um technology for the four versions of the sense amplifier latch [6, 13]: non-scannable latch without data retention, Fig. 6a, scannable latch without data retention, Fig. 6b, non-scannable latch with data retention, Fig. 6c, and scannable latch with data retention, Fig. 6d. The layout for the non-scannable latch without retention fits into 9 tracks, Fig. 7, while the layouts for the scannable latch, and latch with data retention were designed to fit into 12 tracks, Fig. 8 and 9.

For the design in Fig. 9, the retention latch is built out of regular transistors, placed in separate wells for the back biasing capability. This design choice leads to somewhat higher area overhead because of the minimal spacing requirement between wells biased to different potentials. Also, there is an additional area overhead due to well contacts in the retention latch, which can be shared by adjacent latches if they are flipped appropriately. The area overhead for the retention latch that uses high threshold devices is smaller by at least one track in width, but the manufacturing cost is higher. Table 1 gives areas for the four latches in Fig. 6. Although the scannable latch with data retention has four transistors more than the non-scannable latch with data retention, they have the same area, because the height is determined by the height of the retention latch, and the width cannot be reduced because of the minimal well separation ground rule. The scannable latch without data retention has a 2 tracks smaller width than the one with data retention because wells of the transistors in the master and scan/retention latches do not need to be separated.

To determine the exact amount of the power and performance overhead of the scan and data retention mechanisms we developed the layout for multi-bit datapath registers with scan chain and lo-

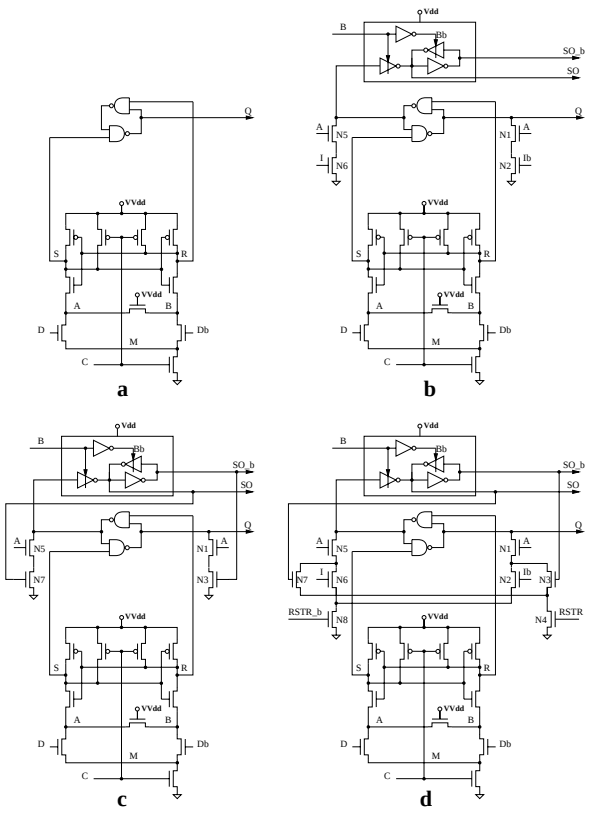

Figure 6: Sense amplifier latch: a - non-scannable latch; b - scannable latch; c - non-scannable latch with data retention; d - scannable latch with data retention.

cal clock distribution network, Fig. 10, and ran simulations on the extracted netlists, according to the methodology described in [13]. Table 1 summarizes results of the simulations.

For measuring delays all latches were loaded with three minimum-size inverters and wiring capacitance. The additional overhead of the data retention feature over the scan mechanism is very small (no more than 1%), and so is the additional overhead of the scan feature over the data retention mechanism. For the implemented designs the increase in delay compared to the non-scannable latch is 15% because very small load capacitances were used in simulations, typical of ultra low power designs, and latch transistors were tuned accordingly. The delay overhead is much smaller in high-performance designs, where transistor sizes are tuned for smaller delays.

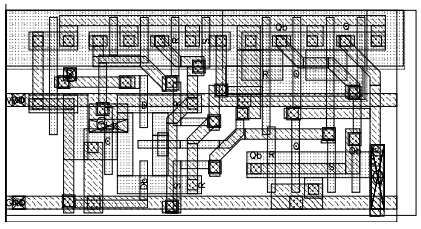

Figure 7: Non-scannable SA latch without data retention, Fig. 6a.

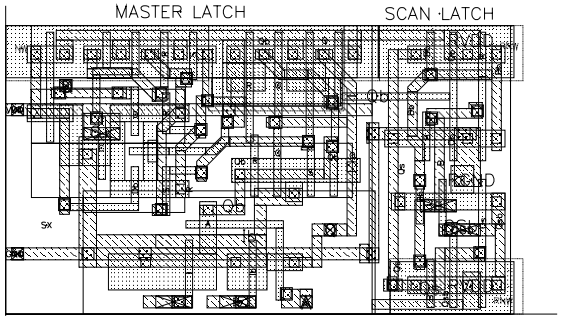

Figure 8: Scannable SA latch without data retention. Fig. 6b.

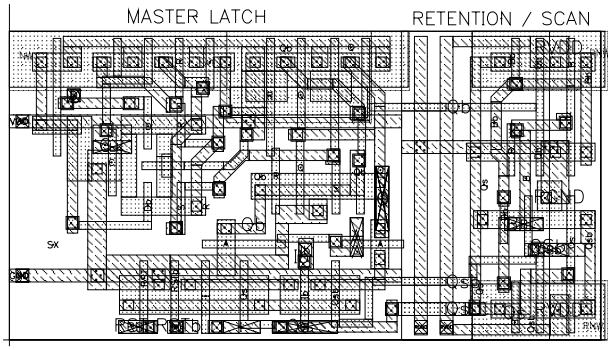

Figure 9: Scannable latch SA latch with data retention. Fig. 6d.

The power of the developed latches was measured using the methodology in [12, 13, 9]. The data in Table 1 correspond to the switching factor $\alpha = 0.3$ and the glitching factor $\beta = 0.16$. The power of the local clock distribution was included into the latch power measurements, and so was the capacitances of metal wires to get from the latch boundary to the data input pin and from the data output pin to the other boundary. A 9-track cell height design was assumed for the non-scannable latch without retention, whereas a 12-track cell height design was used for measuring the power of the other three latches, which resulted in somewhat lower power numbers for the non-scannable latch without retention. If the same bitstep is used for all latches, the power overhead of the scan/retention mechanism is only 3%. The power overhead of the scan/retention mechanism is 3 times smaller than the delay overhead, because the introduced capacitance switches only when new data is latched, which was assumed to happen on the average once every 3.33 cycles ($\alpha = 0.3$). As discussed earlier, the power overhead is even smaller in designs tuned for higher performance.

Figure 10: 16-bit data register (not all layers are shown).

Table 1: Area, delay and average energy per cycle of SA latches with and without scan and data retention mechanisms (Vdd = 0.9V).

latch design	area		delay		energy	
	μ^2	%	ps	%	fJ	%
− scan − retention	23.04	0	174	0	10.63	0
+ scan − retention	40.32	75	198	14	11.13	5
− scan + retention	44.16	92	200	15	11.15	5
+ scan + retention	44.16	92	200	15	11.15	5

4. CONCLUSIONS

A methodology for unifying the scan and data retention mechanismsin latches was presented which leads to scannable latches with the data retention capability achieved at a very low power overhead during the active mode. The exact amount of the power and performance and area overhead was measured using netlists extracted from the layouts built in a state of the art 0.13um technology. The additional overhead of the data retention feature over the scan mechanism is very small, and so is the additional overhead of the scan feature over the data retention mechanism. The proposed integrated scan-retention mechanism achieves both features at the cost of one.

Acknowledgment

The authors would like to thank their colleagues D. Knebel, G. Gristede and A. Haen for the design flow support; K. Warren and J. Moreno for the management support.

5. REFERENCES

[1] F. Klass et al. A new family of semidynamic and dynamic flop-flops with embedded logic for high-performance processors. *IEEE Journal of Solid-State Circuits*, 34(5):712–716, May 1999.

[2] S. Mutoh et al. A 1v multi-threshold voltage CMOS DSP with an efficient power management technique for mobile phone applications. In *ISSCC*, pages 168–169, 1996.

[3] N. Nedovic, M. Aleksic, and V. Oklobdzija. Timing characterization of dual-edge triggered flip-flops. In *ICCD*, August 2001.

[4] N. Nedovic and V. Oklobdzija. Dynamic flip-flop with improved power. In *Proceedings of the International Conference on Computer Design*, September 2000.

[5] N. Nedovic and V. Oklobdzija. Hybrid latch flip-flop with improved power efficiency. In *Proceedings of the Symposium on Integrated Circuits and Systems Design*, 2000.

[6] B. Nikolic et al. Improved sense-amplifier-based flip-flop: Design and measurements. *IEEE Journal of Solid-State Circuits*, 35(6):876–883, June 2000.

[7] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada. A 1-v high-speed MTCMOS circuit scheme for power-down application circuits. *IEEE Journal of Solid-State Circuits*, 32(6):861–869, June 1997.

[8] V. Stojanovic and V. Oklobdzija. Comparative analysis of master-slave latches and flip-flops for high-performance and low-power systems. *IEEE Journal of Solid-State Circuits*, 34(4):536–548, April 1999.

[9] V. Stojanovic, V. Oklobdzija, and R. Bajwa. A unified approach in the analysis of latches and flip-flops for low-power systems. In *Proceedings of the International Symposium on Low Power Electronics and Design*, pages 227–232, August 1998.

[10] J. Tschanz et al. Comparative delay and energy of single edge-triggered and dual edge-triggered pulsed flip-flops for high-performance microprocessors. In *IEEE Symposium on Low Power Electronics and Design*, pages 147–152, August 2001.

[11] C. Webb et al. A 400-MHz S/390 microprocessor. *IEEE Journal of Solid-State Circuits*, 32(11):1665–1675, November 1997.

[12] V. Zyuban and P. Kogge. Application of STD to latch-power estimation. *IEEE Transactions on VLSI Systems*, 7(1), March 1999.

[13] V. Zyuban and D. Meltzer. Clocking strategies and scannable latches for low power applications. In *IEEE Symposium on Low Power Electronics and Design*, pages 346–351, August 2001.