The Impact of Width Granularity on FinFET Latch Operation and Optimization

Huifang Qin, Thomas Ludwig*, Kerry Bernstein**, Jan Rabaey, Edward Nowak***

University of California, Berkeley, *IBM Deutschland Entwicklung, **IBM Thomas J. Watson Research Center, ***IBM Microelectronics

Phone: +1-510-666-3153, email: huifangq@eecs.berkeley.edu

Abstract

This paper investigates the effects of coarse width granularity on latch circuit designs; such granularity may be imposed by FinFET technology. We analyze the sensitivities of delay, power and stability on latch sizing. A performance metric to evaluate Quantization Error (QE) is defined, and used to optimize width granularity in FinFET circuits. Simulation results show that logic latch structure achieves better-optimized performance given its larger sizing compared to that of the General Purpose Register (GPR) latch, while both types of latches operate satisfactorily in a FinFET implementation.

Keywords: FinFET, latch, granularity, quantization

Introduction

The FinFET [1,2] has been considered a forefront candidate for ~10-nm gate length device technology due to its superior scalability, process flow [3], and layout similar to that of conventional MOSFET technology. Functional FinFET inverters [4], SRAMS [5], and ring oscillators [6] have been fabricated, and FinFET IC production may fall as early as the 65-nm technology node (25-nm physical gate length) [7]. While FinFET brings many advantages, its width quantization effect could present a major technology disrupter for FinFET circuit design [8]. In digital design, delay and stability of latch circuits are most sensitive to sizing effects. Thus analysis of FinFET width quantization effects on latch circuits is of great interest for successful digital design migration to FinFET technology.

The FinFET Structure

Fig. 1 illustrates the FinFET structure. A gate straddles a thin, fin-shaped body, forming two self-aligned channels along the fin sidewalls. With two gates controlling the thin channel, short-channel-effects are efficiently suppressed. The effective gate width of a single-fin FinFET is given by two times the fin height ($H_{fin}$). To realize different device widths, multiple fins are used. Thus the resulting quantized fin gate width becomes

$$W = 2 \cdot H_{fin} \cdot N_{fins} = 0.14 \mu m \cdot N_{fins}$$

(1)

for the technology under discussion. $H_{fin}$ in this formula represents the fin height, and $N_{fins}$ is the number of fins in one FinFET device. With this width quantization effect, flexibility for FinFET circuit optimization is limited.

Analysis Setup

Simulations were based on a FinFET 90nm technology model with process corners [9]. As shown in Fig. 2, two types of latch circuits were analyzed, the logic and General Purpose Register (GPR) latches. These two structures represent performance-optimized and area-constrained latch designs, respectively. The logic latch is usually sized larger and operates in open-loop write mode, while the GPR latch is smaller and exhibits higher sensitivity to gate sizing due to its closed-loop write operation. Using the FinFET implementation, power, speed and stability metrics are evaluated for each of these two latches with varied gate sizing. To quantitatively evaluate the optimization of quantized sizing, the fin gate width Quantization Error (QE) at a given performance point is defined as half the maximum distance to the adjacent performance point divided by its performance value. The derivation of QE is illustrated in Fig. 3.

$$QE_i = \max(d_i, d_{i+1}, d_i, d_{i+1})/(2 \cdot D_{max})$$

(2)

where $D_i$ is a performance parameter (delay in Fig. 3) at the i th quantized sizing, and $d_{i-1}$ stands for the distance in that performance parameter to adjacent points. Note that QE is a measure of maximum deviation of performance (delay or signal margin) from that which could be attained with continuously variable FinFET widths.

Logic Latch and GPR Latch Sizing Analysis

To study the sizing effect on latch performance, sensitivity of four performance metrics (delay, active power, clock and data noise margins) to quantized FinFET device sizing were evaluated. The Noise Margin (NM) analysis measures the minimum static disturbance at the clock or data input that causes the latch state to flip. There are three variables in logic latch sizing optimization: width of the input transmission gate (T-Gate) (1), PMOS (2) and NMOS (3) devices in the data-holding inverter loop (see Fig. 2). In this analysis we fixed the

![Figure 1. FinFET structure](image)

![Figure 2. Schematics of logic latch (a) and GPR latch (b).](image)

![Figure 3. Derivation of logic latch delay sizing QE.](image)
size of the NMOS data holding devices, and varied the other two transistor widths with tuning margins of \(\pm (30\% \sim 70\%)\) times the original design value. This range of variation is intended to represent limitations imposed by the requirement that the FinFET physical design fit within the footprint of an original planar design. In the comparisons, sizing of both latches are varied in the same range. Fig. 4 shows the power-speed-stability tradeoff results of a logic latch. The Power-Delay-Product (PDP) metric in Fig. 3(a) is evaluated with delay and active power data. Observed from this plot, PDP of the logic latch can be minimized by optimizing the T-Gate size for a given PMOS size, while other dimensions of performance metrics (PDP at fixed T-Gate width and stability) are monotonically affected by width. The cross points on the curve grids represent data points at discrete FinFET device widths (i.e. physically realizable FinFET designs). The effects of granularity on the logic latch are negligible, with QE for all performance metrics within 3% of the original design value.

Fig. 5 shows the sizing effect on delay, PDP and clock NM of a GPR latch. Compared to the logic latch, the GPR latch is slower due to its smaller size and closed-loop write operation. The compact design and closed-loop writing of the GPR latch results in much greater sensitivity to fin number tuning. QE for GPR latch delay reaches 8.2% when the passgate is made very weak. Thus for low-power designs, a FinFET GPR latch design may deviate from the desired performance by 8.2% at maximum in the given sizing range. In such a case a designer may either relax the timing specification by 8.2%, or design an 8.2% faster GPR latch at the expense of other performance metrics.

From the previous results it can be observed that the quantized T-Gate sizing usually contributes larger error than the PMOS width sizing. A closer look at this one-dimensional granularity is given in Fig. 6. The logic latch, with twice the number of fins as in the GPR latch, achieves about two times better granularity in PDP and delay optimization. Despite its narrower granularity, the FinFET GPR latch can also attain satisfactory performance. Considering only the T-Gate width quantization effect, the closest quantized PDP approximates the optimum point with less than 2% error.

**Conclusion**

Width quantization constraints result in minor degradation to the power-speed-stability optimization of latch circuits with respect to an unconstrained design. While simulation results show that both types of latch can be sized for reasonable performance, the logic latch structure with larger device sizes achieves finer granularity, with QE less than 3%, while the GPR latch may suffer QE as high as 8%. On the other hand, even the

---

**References**


**Figure 4. Logic latch power-speed-stability sensitivities on sizing:** a) Power-Delay-Product (PDP) b) Noise Margin (NM).

**Figure 5. GPR latch sensitivities on sizing:** (a) delay, (b) PDP and (c) clock NM.

8% QE observed in the GPR latch does not translate to an 8% delay penalty, but rather a performance metric representing the worst-case compromise in overpowering this latch to achieve desired performance. Overall this study suggests that the width quantization effect in latches is not a significant barrier to the introduction of FinFET technology.