**Title**
Comprehensive solutions to circuit uncertainty for hardware machine learning system



**Author**
Liu, Chun-chen

**Publication Date**
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Comprehensive Solutions to Circuit Uncertainty for Hardware Machine Learning System

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical Engineering

by

ChunChen Liu

2017

ABSTRACT OF THE DISSERTATION

Comprehensive Solutions to Circuit Uncertainty

for Hardware Machine Learning System

by

ChunChen Liu

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2017

Professor Mau-Chung Frank Chang, Chair

With the coming era of Big Data, hardware implementation of machine learning has become attractive for many applications, such as real-time object recognition and face recognition. The robustness of memory access and monitoring of the circuit uncertainty determines the performance of this kind of application. As current nanotechnology semiconductor device scales down dramatically with additional strain engineering for device enhancement, the overall device characteristic is no longer dominated by device size but also circuit layout and interconnect. A set of test structures was developed to study the timing performance (i.e. propagation delay and crosstalk) of various interconnect configurations and accurately measure and model interconnect parasitic in order to predict interconnect performance on silicon. To improve the validation of the higher order layout effects (WPE, OSE and PSE), Design for Manufacturability (DFM) impacts of two analog layout structures, guard ring and dummy fills impact were monitored with current mirror test circuit using TSMC 28nm HPM process. Finally, a new design of high-speed SRAM with fast access time (cycle time: *650 ps*, access time: *350 ps*), low sensitivity to temperature variation and high reconfigurability (less than 10%

performance difference between 125_rcw_tt vs 0_rcw_tt) was developed. These results described in this thesis provide comprehensive solutions to the requirement of the circuit uncertainty monitoring and rapid memory access of machine learning hardware implementations.

The dissertation of ChunChen Liu is approved.

Jason C S Woo

Wentai Liu

Pei-Yu Chiou

Mau-Chung Frank Chang, Committee Chair

University of California, Los Angeles

2017

# Table of contents

# Figure list

# Table list

# Equation list

# VITA

| | |
|---|---|
| 2003 | B.S., Information Science and Technology, Southeast University, China |
| 2010 | M.S., Electrical Engineering University of California, Los Angeles |
| 2016 | Graduate Student Researcher, University of California, Los Angeles |

# Chapter 1 Introduction

## 1.1 Summary

With the coming era of Big Data, learning-based algorithms for precise prediction are widely developed and applied to modern applications in different fields such as Autonomous Car, Face Identification, Speech Recognition, etc.

In hardware implementation of machine learning algorithms, data transition is a critical part in regards to performance and power consumption [1]. Memory access time can dominate the overall run time of algorithm and therefore reliable SRAM with fast access (finish memory access within one processor clock cycle) is desirable for high-performance machine learning implementation. Unfortunately, conventional SRAM designs cannot achieve such a stringent requirement. As shown in the recent studies[2, 3], conventional SRAM design can only perform stable read/write speed to 0.4~0.8 GHz within normal operating conditions. On the other hand, processor clock can run at > 2GHz. Consequently, it usually takes 3~5 clock cycles to finish memory access operation, which can prompt a bottleneck for hardware Deep Learning (DL) performance.

Advanced package techniques [4, 5] can contribute to the DL memory system performance by improved routing. However, these kinds of solutions tend to be relatively more expensive. Moreover, with the shrink of the CMOS size, the influence of process, voltage, and temperature (PVT) variations become dramatic while conventional SRAM designs only have limited ability to tolerate PVT. When temperatures change, the required read/write operation time for memory may vary as much as 40% [6, 7]. Conventional memory system design has to spare enough margins for worst-case temperature variation (condition), even though most of the memory actually runs at low temperature and can operate faster.

To address the above issues, we propose a high-speed, temperature-variation-immune SRAM compatible with standard digital CMOS technology in this thesis. Compared to state-of-the-art design [8], the proposed SRAM outperforms at faster speed (*2.2GHz vs 800MHz*), size (*121X43 um2 vs 127x44 um2*) and has better temperature variation immunity. Furthermore, 6T Cell with custom design in this proposed SRAM has a smaller area compared to other SRAM design.

The contributions of this thesis are as follows:

1. We designed a high-speed SRAM running at 2.2 GHz. With fast access time, the proposed SRAM can help accelerate hardware in machine learning applications.
2. To improve the validation of the higher order layout effects (WPE, OSE and PSE), Design for Manufacturability (DFM) impacts of two analog layout structures, guard ring and dummy fills impact were monitored with current mirror test circuit using TSMC 28nm HPM process.
3. To accurately measure and model interconnect parasitic in order to predict interconnect performance on silicon, a set of test structures that can be used to study the timing performance (i.e. propagation delay and crosstalk) of various interconnect configurations was developed.
4. We proposed a temperature-variation-immune SRAM design. The proposed work is compatible with conventional SRAM process without additional mask cost. The temperature-variation-immune SRAM design can benefit machine-learning applications, since intensive memory access of machine learning can significantly vary due to the temperature of SRAM.
5. The smaller size of bank is designed to gain higher reconfigurability and faster read/write. The machine-learning algorithm could turn off some neurons to avoid unnecessary leakage current with less power consumption.

## 1.2 Thesis overview

In Chapter 1, we provide a summary of the contributions made and the chapter dissertation. In subsequent chapters, we describe in detail the design of high-speed SRAM and solutions developed to detect the circuit uncertainty, along with the impacts of these solutions to the advance of hardware machine learning application.

In Chapter 2, we begin by presenting a review of current hardware machine learning application. In Chapter 3, we briefly describe the high order layout effects and the solutions we developed are detailed in Chapter 4.

In Chapter 5, we develop an interconnect test structure for RC parasitic validation, to accurately measure and model interconnect parasitic in order to predict interconnect performance on silicon.

In Chapter 6, we detail the proposed temperature-variation-immune SRAM design for the memory access in machine learning hardware application.

In Chapter 7, we review the contributions of the thesis, and list possible directions of future research.

# Chapter 2 Hardware machine learning applications

With the advent of big-data era, machine learning has become increasingly powerful in solving problems from various domains such as face recognition in security screening and high-frequency trade in banking. However, most machine learning algorithms are complex in nature and limited to real-time operation with software implementation. Accordingly, the hardware accelerated or learning on-chip [9-13] are evolved, making use of powerful heterogeneous hardware platforms involving graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and/or network-on-chips (NoCs). For example, Intel last year released Xeon processors with built-in FPGAs dedicated for data center and learning applications [14].

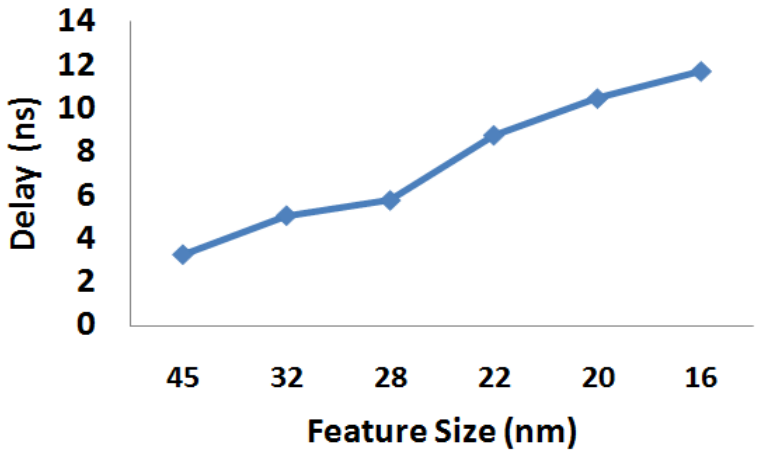


Figure 1. Delay of 1-mm global interconnect in different technology nodes based on data reported in International Technology Roadmap for Semiconductors (ITRS) [15]

A major bottleneck in these heterogeneous platforms lies in interconnects between various system components; as will be demonstrated in Section 2, the data latency is the decisive factor of the overall performance. On the other hand, with relentless technology scaling, the features size is shrunk rapidly and the wire width and spacing are all significantly reduced, resulting in

high coupling capacitance. To minimize the crosstalk impact, the wire thickness is also reduced with the major drawback of high sheet resistance due to a smaller cross-section area. The high sheet resistance leads to serious issue for Power Distribution Network (PDN) and clock tree design. Figure 1 illustrates that the interconnect delay increases drastically with technology scaling. The delay is increased several times from CMOS 45nm to 16nm process and it becomes more serious for 10nm and below.

Accordingly, it is crucial to develop an accurate interconnect model for accurate system performance simulation and prediction. Currently, many test structures [16-20] have been developed to characterize the impacts of standard cell architecture, custom data path, current mirror and I/O pad design. However, only a few test structures exist with a focus on the impact of interconnect. In addition, in order to be useful and practical, they must be calibrated with measured data, to establish silicon-to-model correlation [21]. The conventional cross-bridge Kelvin structure [22, 23] needs extra probe pads to directly measure the on-chip parasitic effect, which is not suitable for interconnect monitoring in real chip designs. Then, simple ring oscillator [24, 25] is developed to measure the frequency of various interconnect configurations. This approach is widely adopted by the foundries and fabless design houses to validate the interconnect performance. However, it mainly focuses on the single interconnect and ignores the effect of cross coupling between adjacent wire impacts toward overall performance. Also, on-chip interconnect monitoring based on time-to-digital converter (TDC) has been proposed [26], but it suffers from the non-idealities of TDC. A compact yet comprehensive test structure to capture all interconnect parasitic is still a missing piece in the literature.

To address this issue, this thesis describes a test structure based on a set of enhanced ring oscillator designs. It not only measures the propagation delay of various interconnect structures but also accounts for different crosstalk impacts (i.e. in-phase and out-of-phase crosstalk).

Compared with current interconnect test structure, this proposal can measure both interconnect delay and crosstalk impact at the same time to significantly reduce the test structure area. It is easy to embed in the real chip for interconnect validation during the production. A first-order empirical model is also put forward to estimate the RC parasitic for silicon-to-model correlation. Since the test structure is relatively simple and small, it is easy to implement in real chips to monitor actual RC parasitic variations during manufacturing. We have validated the proposed structure on a TSMC 28nm test chip, and showed its efficacy on our in-house Neuro Processing Unit (NPU).

# Chapter 3 High order layout effects

With the advances in nanotechnology, the semiconductor device is scaled down rapidly with additional strain engineering for device enhancement. The CMOS performance is improving [39-46] in various applications, such as RF [47-54,71,72], wireline SerDes [55-60, 65,68-70], analog signal processing [61-64], and digital signal processing [12-14]. However, the overall device characteristic is no longer dominated by the device size but also layout effects (WPE, OSE and PSE) [27]. It is critical to understand Design for Manufacturability (DFM) impacts with various layout topology toward the overall circuit performance. Currently, the digital standard cell and analog differential pair layout test structure are implemented to validate the layout effects. However, two important analog circuit layout topology - guard ring and dummy fill impact - are not well studied yet. Therefore, this thesis describes a current mirror test circuit to examine the guard ring and dummy fills DFM impacts using TSMC 28nm HPM process.

For analog design, the circuit performance is highly dependent on the device matching; the centroid topology is often chosen to minimize the layout environmental variation. The transistors are arranged symmetrically in centroid style where all the devices suffer from the same physical and electrical impacts from all directions. The centroid topology focuses the active devices layout impacts only; the guard ring protection and dummy fill layout structures are not fully taken into consideration yet. At a result, the modified current mirror configurations are implemented to explore various guard ring and dummy fill DFM impacts.

In the next chapter, we describe two analog layout structures: the guard ring and dummy fill impact that were not well studied yet. Then, we illustrate the current mirror test circuit designed to examine the guard ring and dummy fills DFM impacts using TSMC 28nm HPM process. At the end of the next chapter, the measurement results have been taken into consideration for TSMC 7nm FF enhanced DFM guideline for analog circuit yield enhancement.

# Chapter 4 Analog circuit guard ring and dummy fill DFM analysis

## 4.1 Current Mirror Design

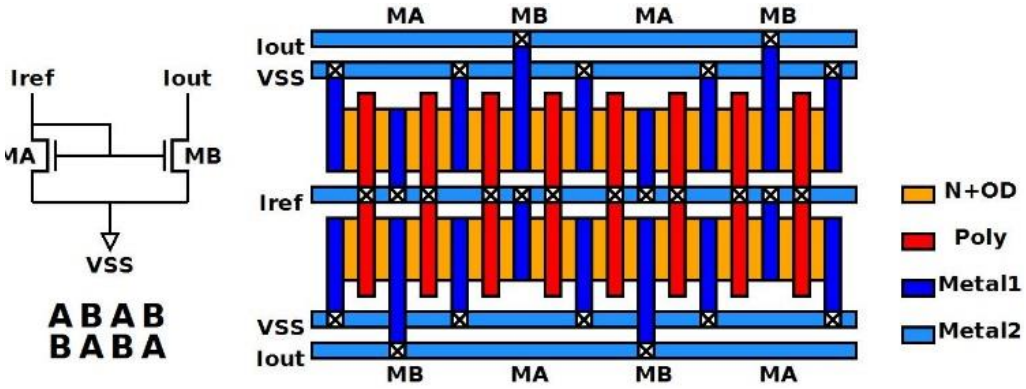


Figure 2. NMOS Current Mirror Fine Centroid Topology

Figure 2 shows the conventional current mirror [28] centroid layout topology. The multi-finger devices MA and MB are placed alternately and symmetrically. The individual transistors experience the same layout impacts, suffering the same physical and electrical impacts from all direction [29]. This layout topology was originally developed to minimize the angular implant doping variation. It is further enforced to reduce layout effect (WPE, OSE and PSE) impacts since 45nm process [30, 31]. In this chapter, we focus on the fine centroid layout style rather than the coarse one where a group of transistors are arranged symmetrically to minimize high interconnect RC parasitic impacts.

## 4.2 Guard Ring

In order to isolate the current mirror from other impacts, the active devices are protected by guard ring and diffusion (OD)/poly (PO) dummy. The guard ring is typically used to protect the active devices from latch-up and noise interference where P+ guard ring with VSS

connection protects NMOS active devices, PMOS active devices surrounds with N+ guard ring connected to VDD, as shown in Figure 3. There are two kinds of guard ring: single and double guard ring where the single guard ring employs either P+ or N+ guard ring only and the double guard ring that mixes with both P+ and N+ guard ring.

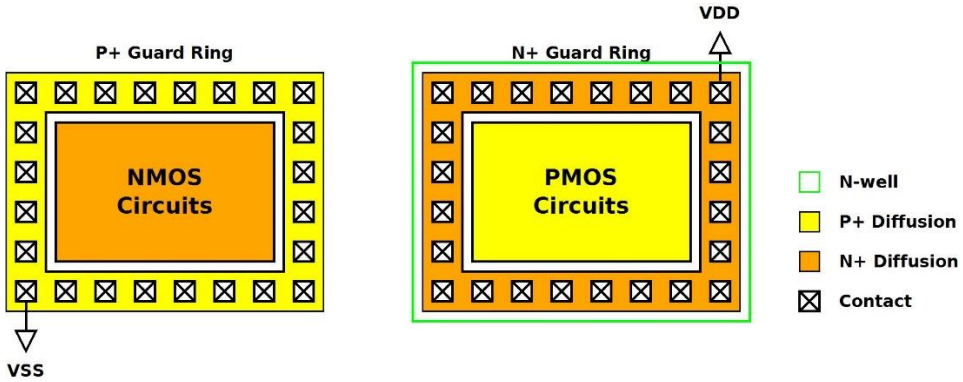


Figure 3. P+/N+ Guard Ring

Currently, the simulation model only considers the active device layout effect impacts; it ignores the physical and electrical impacts introduced by guard ring. Only diffusion spacing between active devices and guard ring are considered in simulation using OSE model. The guard ring diffusion width and P+/N+ implant type both contribute to the device mobility changes. Therefore, we propose to modify the original OSE model [32] by introducing effective STI width ($STIW_{eff}$) parameter. As a first-order model, we set a threshold of OD width of single guard ring ($ODW_{th}$) when guard ring effect comes into play. The value of $ODW_{th}$ can be found by experiment. If ODW is smaller than $ODW_{th}$, then the $STIW_{eff}$ is defined as

Equation 1. Effective STI width ($STIW_{eff}$) parameter

$$STIW_{eff} = STIW \times \left(1 + K\frac{ODW_{th}}{ODW}\right)$$

Where K is curve fitting parameter.

## 4.3 Dummy Fill Structure

Dummy fill is typically related with three different types of dummy: diffusion, poly and metal. This section mainly focuses on diffusion (OD) dummy fill because the diffusion uniformity is directly related with Shallow Trench Isolation (STI) and Rapid Temperature Annealing (RTA) process that are linked with the transistors formation. The poly (PO) dummy fill is critical for 3D FinFet technology because the height of FinFet is directly dependent on the PO dummy density. Finally, the metal dummy is closely linked to Chemical Mechanical Polishing (CMP) process and directly impacts toward interconnect RC parasitic. For current test chip implementation, it is divided into two-level dummy fill. The first level dummy structure is similar to current mirror with same dimension and diffusion type; the second level dummy structure is used to examine different OD fill impacts, as shown in Figure 4.

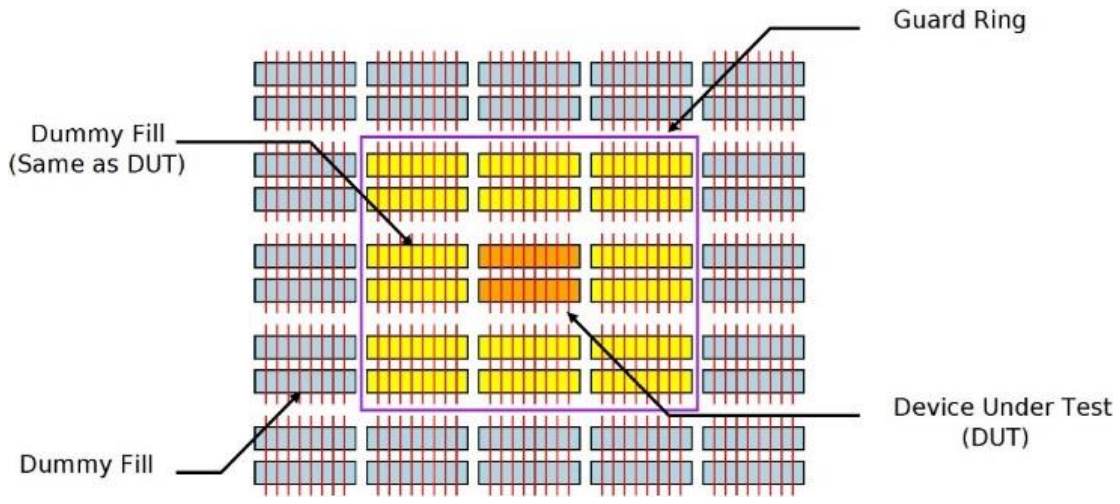


Figure 4. Device under Test (DUT) Structure

Most of OD dummy fill is limited to N+OD dummy fill in order to simplify the chip implementation; it further explores with P+OD and mixed N+OD/P+OD dummy fill impacts toward different current mirror configurations. The conventional fine centroid PMOS/NMOS current mirrors are chosen with N+OD dummy fills as reference circuit for comparison because N+OD dummy fill is commonly offered by foundry. It is easy to implement compared with P+OD and mixed one. For test chip implementation, it focuses on the different type of diffusion impacts toward current mirror performance. All the poly, diffusion and metal density are same for all test structures, except for the type of diffusion. Three types of diffusion, N+OD, P+OD and(?) mixed on,e are implemented. The type of diffusion is highly related with Rapid Temperature Annealing (RTA) process for device threshold voltage and leakage variation [33]. The effect threshold voltage can be expressed as:

Equation 2. The effect threshold voltage

$$V_{teff} = V_{tref}\left(1 + f(D_{NOD}, D_{POD})\right)$$

where $D_{NOD}$ and $D_{POD}$ are dummy density for N+OD and P+OD within 100 μm x 100 μm window around MOSFETs, respectively, and $f(D_{NOD}, D_{POD})$ is a look-up table function which is obtained by measurement. Therefore, the test structure is used to study threshold voltage changes with different types of dummy diffusion fills.

## 4.4 Test Chip Results

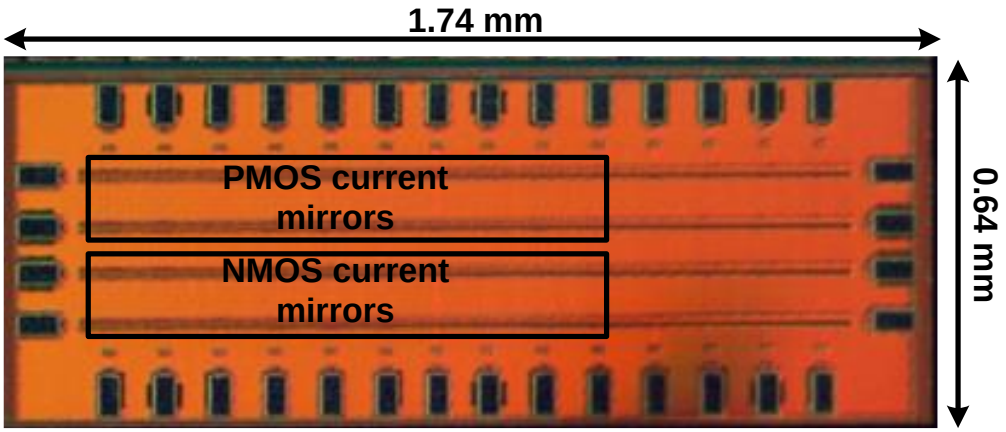


Figure 5. Guard Ring and Dummy Fill Die photo

In order to validate the analog test structure, various current mirrors are implemented using TSMC 28nm HPM process, as shown in Figure 5. [73] It includes five PMOS (Table 1) and five NMOS (Table 2) current mirrors. The distance between guard ring and active devices is 1μm. For single guard ring, the OD widths in guard ring are 0.14 μm (1X) and 0.28 μm (2X), respectively. For double guard ring, the OD width is 0.28 μm for both P-type and N-type guard ring. The distance between N-type and P-type guard ring is 0.4 μm. For OD dummy fill, the total OD density is over 50% within 100 μm x 100 μm window. For N+/P+ OD dummy fill, half of OD dummy is P-type and the other half is N-type.

Table 1. PMOS Current Mirror Measurement

| Type | Guard Ring | Dummy | Simulated Ratio | Measured Ratio |
|---|---|---|---|---|
| PMOS | Double | P+OD | 1.00 | 1.00 |
| PMOS | Single 1X | P+OD | 1.00 | 1.01 |
| PMOS | Single 2X | P+OD | 1.00 | 1.05 |
| PMOS | Double | N+OD | 1.00 | 0.99 |
| PMOS | Double | N+/P+OD | 1.00 | 0.99 |

In measurement, the same R (off-chip potentiometer) is connected to all current mirrors (between MA and VDD for NMOS, MA and VSS for PMOS), and the output current is measured and compared with output current of the reference current mirror (P+OD and double guard ring for PMOS, N+OD and double guard ring for NMOS and N+OD). The listed current ratio is the average of multiple (10) chips.

Table 2. NMOS Current Mirror Measurement

| Type | Guard Ring | Dummy | Simulated Ratio | Measured Ratio |
|---|---|---|---|---|
| NMOS | Double | N+OD | 1.00 | 1.00 |
| NMOS | Single 1X | N+OD | 1.00 | 1.01 |
| NMOS | Single 2X | N+OD | 1.00 | 1.05 |
| NMOS | Double | P+OD | 1.00 | 1.09 |
| NMOS | Double | N+/P+OD | 1.00 | 1.10 |

From Table 1, the single guard ring has 1%~5% larger current compared with double guard ring because the PMOS guard ring introduces additional tensile stress to enhance mobility, and the tensile stress is further increased with wider guard ring (i.e. 2X). This impact is predicted by proposed modified OSE model. The comparison of simulation with original OSE model, with proposed OSE model and measurement is shown in Figure 6. For double guard ring, the P+/N+ guard rings offset each other's(?) performance enhancement. Regard to different OD fill, the impacts are relatively minor with 1% performance difference. This means *f*(D_NOD,D_POD ) has small value for PMOS.

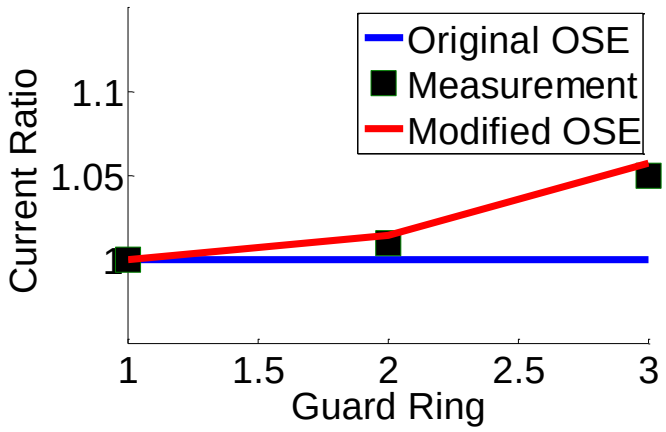


Figure 6. Comparison of original OSE model, measurement and modified OSE model

From Table 2, NMOS current mirror performance is different from PMOS due to compressive stress that degrades the device driving capability. The single guard ring shows the performance enhancement similar to PMOS one. P+OD and mixed OD dummy fills increase the current by about 10%.

# Chapter 5 Interconnect test structure for RC parasitic validation

## 5.1 Interconnect structures

In this section, we briefly review the interconnect structures and models as they directly relate to our proposed test structures.

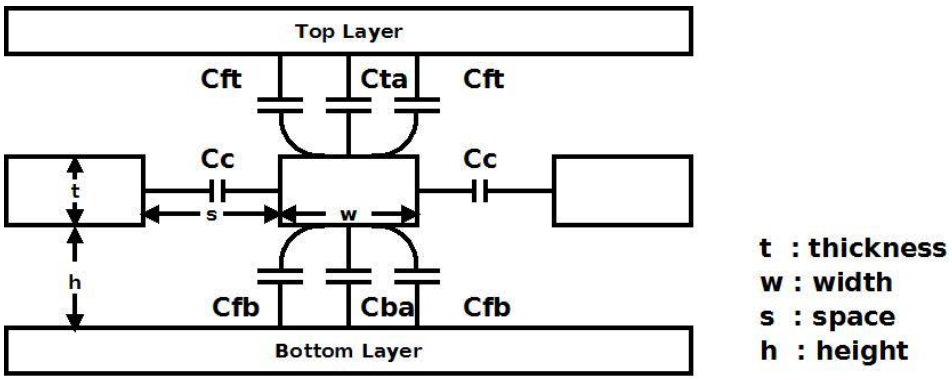


Figure 7. Interconnect model

The interconnect model where the wire is routed through top and bottom layer and coupled with two left/right adjacent wires is shown in Figure 8. The capacitances between the layers are named area capacitance ($C_a$) and fringe capacitance ($C_f$). They are further grouped into top ($C_{top}$) and bottom ($C_{bottom}$) capacitance as follows:

Equation 3. Top capacitance

$$C_{top} = C_{ta} + 2\ C_{ft}$$

Equation 4. Bottom capacitance

$$C_{bottom} = \ C_{ba} + 2 \ C_{fb}$$

Where $C_{ta}$ is top area capacitance, $C_{ba}$ is bottom area capacitance, $C_{ft}$ is top fringe capacitance, and $C_{fb}$ is bottom fringe capacitance.

The total capacitance is defined as the sum of the top, bottom and coupling capacitance ($C_c$) as follows:

Equation 5. Total capacitance

$$C_{total} = C_{top} + \ C_{bottom} + 2C_c$$

When wires run parallel to each other, the signal (victim) propagation is affected by the adjacent wires (aggressors) through the coupling capacitance, as shown in Figure 9. It is called crosstalk impacts.

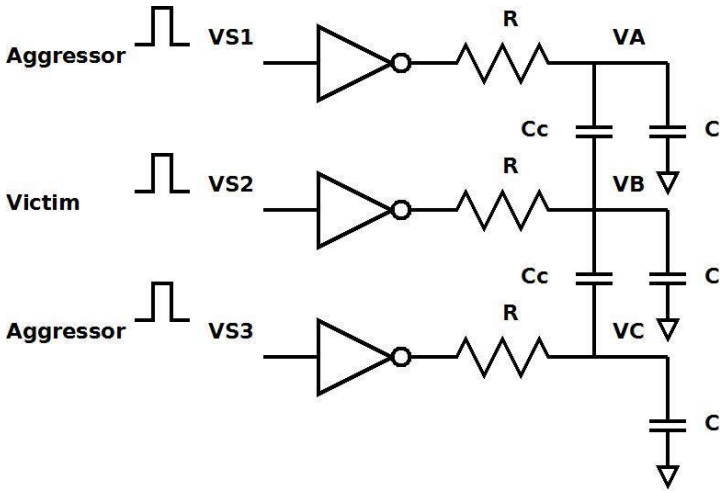


Figure 8. Crosstalk model

The crosstalk is highly dependent on the spacing between the adjacent wires (victim and aggressors) as well as the traveling direction. The effective coupling capacitance is significantly reduced when the spacing is increased. It also changes with different traveling directions due to the Miller effect. If the aggressor remains constant voltage, the victim is said to be quiet or shielded without any crosstalk impact. If the signals in both victim and aggressor travel in the same direction, the crosstalk is smaller (called in-phase crosstalk). If they travel in opposite directions, the effective coupling capacitance is larger (called out-of-phase crosstalk).

Assuming there is no switching in top and bottom routing layers, the left and right aggressor introduces the noise on the victim through the coupling capacitance ($C_c$). The effective capacitance ($C_{eff}$) of the victim can be approximated using the following equations:

Equation 6. Ground capacitance

$$C_{gnd} = C_{top} + C_{bottom}$$

Equation 7. Quiet mode

$$C_{eff} = C_{gnd} + \ 2C_c$$

Equation 8. In-phase crosstalk

$$C_{eff} = C_{gnd}$$

Equation 9. Out-of-phase crosstalk

$$C_{eff} = C_{gnd} + \ 4C_c$$

The crosstalk can be modeled using equivalent lump model in the Laplacian domain [26]. In order to simplify the model, all resistance $R$, capacitance $C$ and coupling $C_C$ are the same for all branches. $V_{S1}$ and $V_{S3}$ are the aggressor input voltage while $V_{S2}$ is the victim voltage. The output voltages $V_A$, $V_B$ and $V_C$ are defined as follows:

Equation 10. output voltage $V_A$

$$V_A = \frac{(1 + a_1 s + a_2 s^2)V_{s1} + (a_3 s + a_4 s^2)V_{S2} + a_5 s^2 V_{s3}}{(1 + b_1 s)(1 + b_2 s)(1 + b_3 s)}$$

Equation 11. Output voltage $V_B$

$$V_B = \frac{a_6 V_{s1} + (1 + a_7 s)V_{s2} + a_8 V_{s3}}{(1 + b_4 s)(1 + b_5 s)}$$

Equation 12. Output voltage $V_C$

$$V_c = \frac{a_5 s^2 V_{s1} + (a_3 s + a_4 s^2)V_{S2} + (1 + a_1 s + a_2 s^2)V_{s3}}{(1 + b_1 s)(1 + b_2 s)(1 + b_3 s)}$$

Where

$a_1 = 2RC + 3RC_c$

$a_2 = R^2C^2 + R^2{C_c}^2 + 3R^2CC_c$

$a_3 = RC_c$

$a_4 = R^2{C_c}^2 + R^2CC_c$

$a_5 = R^2{C_c}^2$

$a_6 = RC_c$

$a_7 = RC + RC_c$

$a_8 = RC_c$

$b_1 = RC$

$b_2 = RC + RC_c$

$b_3 = RC + 3RC_c$

$b_4 = RC$

$b_5 = RC + 3RC_c$

In order to examine different crosstalk operations, $V_B$ is set to be a step function (with Laplacian transform of V/s) and $V_A$, $V_C$ are either set at zero for quiet mode operation or a step function

(with Laplacian transform of ±V/s) with different polarity to model in-phase and out-of-phase crosstalk. It is then transformed to time domain and simplified through first order approximation.

# 5.2 Ring Oscillator Configuration

In this section, we proposed an enhanced test structure derived from ring oscillators. The overall structure is shown in Figure 10. It consists of an input control unit and three sets of ring oscillators. The control unit controls the test signal (i.e., victim) and two left/right adjacent routing signals (i.e., aggressors) for different crosstalk operations (i.e., quiet mode, in-phase and out-of-phase crosstalk). Finally, the test signal is fed into the down counter to scale down the ring oscillator frequency for oscilloscope measurement.

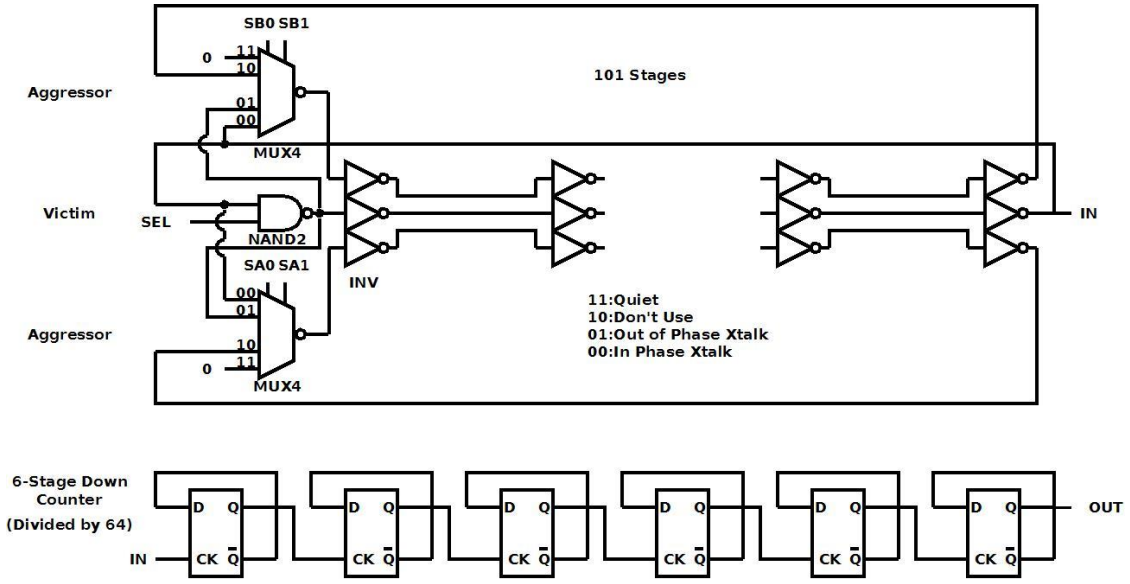


Figure 9. Interconnect Test Structure

The control unit is further divided into NAND2 and 4-inputs MUX; NAND2 is used to control the test signal (victim) propagation [74]. When the control signal SEL is low, the output of NAND2 is always set to high, which results in no oscillation. When the control signal SEL is high, the output of NAND2 is inverted by the input signal IN and the signal propagates

through the inverter chain and creates oscillation.

The aggressors are controlled by two 4-inputs MUX with in-phase crosstalk (00), out-of-phase crosstalk (01), quiet mode (11) and don't use (10) operations. If the select signals SAx and SBx are set to "10", it is defined as "don't use" or invalid state. If they are set to "11", it models the signal propagating along through the inverter chain with shielded protection while the aggressors are set to one or zero to avoid any crosstalk impacts. If they are set to either "00" or "01", the input/output of NAND gates are fed into MUX to toggle the adjacent routing signals. The signals travel in the same or opposite directions as test one for in-phase and out-of-phase crosstalk study. The spacing between the victim and aggressors can be further adjusted to examine the various spacing crosstalk impacts toward signal propagation.

The proposed test structure is not limited to frequency measurement. An empirical model is also developed to estimate RC parasitic. It can correlate silicon measurement with simulation results as a practical way to identify the source of mismatch for improving interconnect mismatch.

## 5.3 Empirical Model

Compared with conventional ring oscillator approaches, the proposed test structure does not only predict the interconnect behavior through frequency measurement; the RC parasitic can be calculated through a set of ring oscillators and correlate with results from silicon measurements. The empirical model is derived as follows:

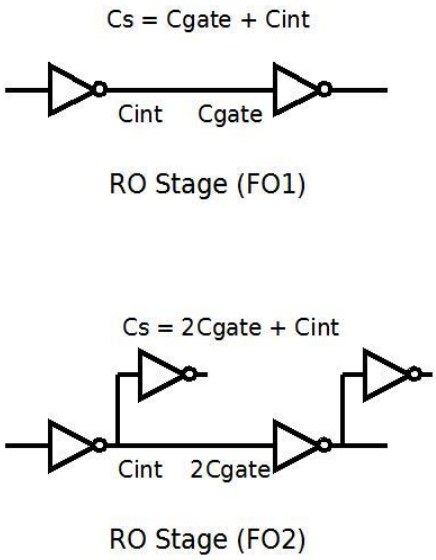


Figure 10. Reference and 2 Fanout RO Stage

Typically, the delay is expressed in terms of supply voltage, average current and load capacitance as

Equation 13. The delay

$$T = C\frac{V}{I}$$

Where it is also rewritten in terms of PMOS/NMOS saturation current

Equation 14. The alternative delay

$$T = CV\left(\frac{1}{I_{dp}} + \frac{1}{I_{dn}}\right)$$

Where $I_{dp}$ is PMOS $I_{dsat}$ and $I_{dn}$ is NMOS $I_{dsat}$ .

Each stage switches twice during the complete cycle and the ring oscillator delay is calculated as follows:

Equation 15. Ring oscillator output frequency

$$f_d = \frac{1}{2nT_s}$$

Where $f_d$ is the ring oscillator output frequency, n is the number of inverter stages and $T_s$ is the stage delay.

The ring oscillator delay $T_{osc}$ with output down counter scaling factor m is defined as

Equation 16. Ring oscillator delay

$$T_{osc} = mT_d$$

Equation 17. Down counter scaling factor

$$T_{osc} = 2nmT_s$$

The stage capacitance $C_s$ can be estimated using Equation 13.

Equation 18. Stage capacitance

$$C_s = \frac{T_s I_{eff}}{V_{dd}}$$

Where $I_{eff}$ is the effective current and is defined as the difference between the active current ($I_{dda}$) and leakage current ($I_{ddq}$) (i.e. $I_{eff} = I_{dda} - I_{ddq}$).

If $I_{ddq}$ is quite small, it is ignored during the calculation; then Equation 18 is rewritten as

Equation 19. Stage capacitance

$$C_s = \frac{T_{osc} I_{eff}}{2mnV_{dd}}$$

From Equation 19, the stage capacitance is calculated using in-phase crosstalk delay and current measurement rather than the shielded one because it eliminates coupling capacitance impacts.

The time delay is calculated with the switching resistance $R_{sw}$ and stage capacitance $C_S$ by

Equation 20. Time delay

$$T_s = R_{sw} C_s$$

It can be simplified as

Equation 21. switching resistance

$$R_{sw} = \frac{V_{dd}}{2\ I_{eff}}$$

Since the crosstalk introduces voltage noise on the victim, the overall delay and current measurement is changed; then, the quiet one measurement is chosen for switching resistance calculation.

The input gate capacitance $C_{gate}$ and output interconnect capacitance $C_{int}$ can be further estimated using two sets of ring oscillators with single fanout (FO1) and double fanout (FO2).

The stage capacitance $C_s$ can be divided into input $C_{in}$ and output one $C_{out}$ shown in Figure 10:

Equation 22. Capacitance correlation

$$\frac{T_{osc1}}{2mn} = R_{sw}\left(C_{int} + C_{gate}\right)$$

Equation 23. Capacitance correlation

$$\frac{T_{osc2}}{2mn} = R_{sw}\left(C_{int} + 2\ C_{gate}\right)$$

From Equation 22 and Equation 23, we can derive the output capacitance and input capacitance as

Equation 24. Output capacitance

$$C_{gate} = \frac{T_{osc2} - T_{osc1}}{2mnR_{sw}}$$

Equation 25. Input capacitance

$$C_{int} = \frac{2\ T_{osc1} - T_{osc2}}{2mnR_{sw}}$$

With aggressor zero or step inputs, the victim output is evaluated through inverse Laplace Transform and expressed in Equation 26 (in-phase), Equation 27 (out-of-phase) and Equation 28 (quiet mode) crosstalk operation:

Equation 26. In-phase victim output

$$V_i(t) = \left(1 - e^{-\frac{t}{RC}}\right)V_{dd}$$

Equation 27. Out-of-phase victim output

$$V_o(t) = \left(1 + \frac{2}{3}e^{-\frac{t}{RC}} - \frac{2}{3}e^{-\frac{t}{R(C+3C_c)}}\right)V_{dd}$$

Equation 28. Quiet mode victim output

$$V_q(t) = \left(1 - \frac{1}{3}e^{-\frac{t}{RC}} - \frac{2}{3}e^{-\frac{t}{R(C+3C_c)}}\right)V_{dd}$$

Equation 26-28 are further simplified using the Taylor series expansion. Since RC constant is quite small, the second and higher coefficients are ignored for first-order approximation. Moreover, the coupling capacitance $C_c$ is much higher than the capacitance C (sum of input gate capacitance $C_{gate}$ and interconnect top/bottom capacitance: $C_{top}$/$C_{bottom}$). The term $R$(C+3$C_c$) is rewritten as 3$RC_c$. Moreover, the scaling factor ½ is taken into consideration to convert the lump model into a distributed one in order to match the simulation results. Finally,

the capacitance $C$ and $C_c$ are calculated as follows:

Equation 29. Capacitance

$$C = \frac{T_o T_q}{R(T_o + T_q)}$$

Equation 30. Coupling capacitance

$$C_c = \frac{2T_o T_q}{3R(2T_o - T_q)}$$

Where $T_o$ is the out-of-phase time delay and $T_q$ is the quiet one.

The empirical model can estimate the first order interconnect RC parasitic through simple measurements from our test structure. It is useful to monitor in-die and die-to-die RC parasitic variation and provide feedback to identify the source of variation during real chip production.

## 5.4 Test Structure Characterization

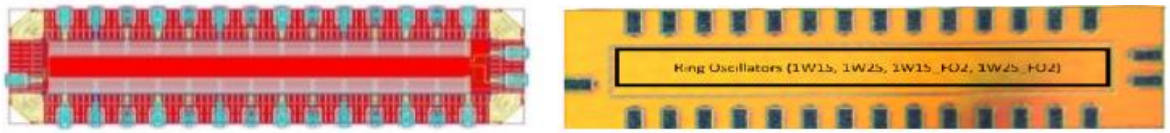


Figure 11. Ring Oscillator Design

In order to validate the test structures, four ring oscillators are implemented and taped out using TSMC 28nm HPM process, as shown in Figure 12. They are divided into single width single spacing (1W1S) and single width double spacing (1W2S) ring oscillator with single & double fanout (FO1 & FO2) where single width and single spacing are referred to minimum line width and spacing. Every ring oscillator consists of control units and 100 inverter stages, and every inverter drives 50um long Metal3 wire. The low Vt clock inverter is employed to minimize the insertion delay impacts and the large number of inverter stages are used to minimize the local device process variation. The 50um long Metal3 wire is chosen to fit the test chip area requirement with typical metal fill to achieve the density requirement.

For typical interconnect process, it is divided into three different metal configurations: the thin metal layer (Mx), the intermediate metal layer (My) and thick metal layer (Mz). The thin metal layer is often utilized for standard cell design and local routing (< 20um). The intermediate metal is targeted for global routing (> 20um), especially for clock tree design. The thick metal layer is used for Power Distribution Network (PDN) to minimize IR drop. Therefore, the intermediate metal layer (i.e. Metal3) is chosen for test chip implementation.

The current test chip structure has been modified for TSMC 7nm FF interconnect process evaluation; the actual wire width, spacing and length are revised to meet the transition time, duty cycle, crosstalk, jitter and electromigration requirements. The metal density variation is also taken into consideration due to high interlayer coupling as well as Chemical-Mechanical Planarization (CMP) effect. For TSMC 7nm FF process. The overall system performance is not only dominated by the device but also interconnect. Moreover, the test structures are used as figures of merit to compare the interconnect performance among different foundries. Based on preliminary simulation with internal enhanced interconnect model, the modified test structures have identified a few TSMC 7nm FF interconnect issues and have been incorporated into critical path design.

Figure 13 shows that the output waveforms of various test structures: quiet, in-phase and out-of-phase. The out-of-phase delay is longer, followed by quiet delay, then the in-phase.

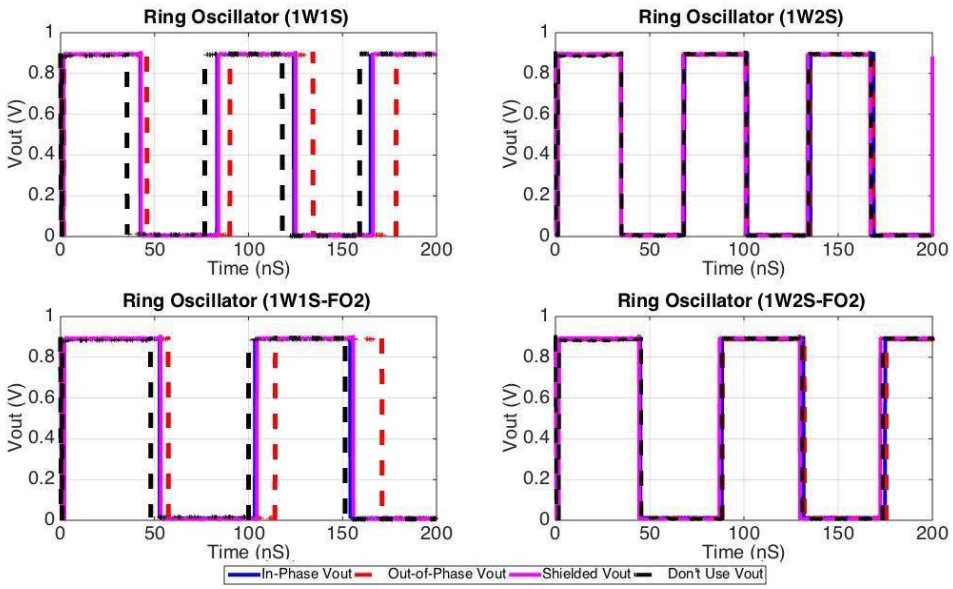


Figure 12. Ring oscillator waveforms with different configuration

The actual stage delay is highly related to the wire spacing. For single width single spacing configurations, the delay is improved with in-phase and out-of-phase crosstalk. With the double spacing, it reduces the coupling capacitance; the overall delay is similar among the three test structures. The measured ring oscillator time period ($T_{osc}$) and current ($I_{eff}$) results are shown in Table 3. The current ($I_{eff}$) is the difference between the normal current ($I_{dda}$) and the leakage one ($I_{ddq}$)

Table 3. Ring Oscillator Simulation Results

| Ring Oscillator | In-Phase Mode | | Out-of-Phase Mode | | Quiet Mode | |
|---|---|---|---|---|---|---|
| | Tosc (ns) | Ieff (uA) | Tosc (ns) | Ieff (uA) | Tosc (ns) | Ieff (uA) |
| 1W1S (FO1) | 81.66 | 891.50 | 88.39 | 990.63 | 82.31 | 503.47 |
| 1W1S (FO2) | 101.35 | 1388.00 | 113.22 | 1384.87 | 102.43 | 778.20 |
| 1W2S (FO1) | 65.99 | 1079.07 | 66.87 | 1093.13 | 66.22 | 532.07 |
| 1W2S (FO2) | 86.5 | 1518.57 | 87.00 | 1534.97 | 85.32 | 817.23 |

Through various ring oscillator structures, the RC parasitic can be calculated as shown in Table 4. We compare the data measured from our test structure with that from the state-of-art test structure [34]. Note that [34] can only measure quiet mode due to the lack of cross-talk aggressors. State-of-art test structure [34] estimates the total stage capacitance and switching resistance and cannot estimate $C_{int}$, $C_c$ and $C_{gate}$. Using new test structures (i.e. in-phase and out-of-phase crosstalk), $C_{int}$, $C_c$, $C_{gate}$, $C_{total}$ and $R_{sw}$ can be calculated.

Table 4. Model/Specification Comparisons

| Parameter | 1W1S | | | 1W2S | | |
|---|---|---|---|---|---|---|
| | this work | [34] | Spec | this work | [34] | Spec |
| Ctotal (fF) | 12.51 | 14.24 | 12.39 | 12.24 | 12.12 | 10.68 |
| Cgate (fF) | 3.02 | N/A | 2.54 | 3.82 | N/A | 2.54 |
| Cint (fF) | 9.50 | N/A | 9.85 | 8.42 | N/A | 8.14 |
| Cc(fF) | 6.82 | N/A | 7.91 | 6.81 | N/A | 5.51 |
| Rsw (Ω) | 504 | 497 | 450 | 417 | 423 | 276 |

From the above tables, it is shown that our empirical model prediction is close to the design specification using various crosstalk modes. There are 1% (1W1S) and 15% (1W2S) Ctotal errors as well as 12% (1W1S) and 51% (1W2S) $R_{sw}$ errors between measurement and specification. The larger error of capacitance estimation in 1W2S is due to the fact that oscillation period in 1W2S is smaller than in 1W1S, and relative measurement error for oscillation period increases in 1W2S. Moreover, the large error may be related with double spacing OPC and CMP operation resulting in line width, spacing and thickness changes in actual silicon fabrication. With approach in [34], $C_{total}$ errors are around 14% and $R_{sw}$ errors are similar with proposed model. The total delay estimation errors of proposed test structure are 12% and 73% for 1W1S and 1W2S cases, while errors of [34] are 27% and 74% for 1W1S and 1W2S case. The results are further emphasized and (?) the proposed empirical model has a significant improvement toward the state-of-art test structure especially in small wire spacing case.

Due to limited number of the samples available, the current test chip is targeted for parasitic validation only; however, the test structure can bridge the gap between design and process for parasitic evaluation. It can be further implemented in test vehicle or real chip (i.e. scribe line) to measure the interconnect process variation and identify potential process failure mechanism.

## 5.5 System Impact

To further validate the efficacy of the proposed test structure in real designs, we apply it as an interconnect monitor to our in-house NPU taped out with 28 nm process. The NPUs implements AlexNet [35] for image recognition. The chip area is 9mm x 9mm. It has 7 hidden layers, 650,000 neurons and 60,000,000 weight parameters. The delay profile of interconnect can vary significantly among chips located in different corners of the wafer. The die photo is shown in Figure 13. As can be seen from the figure, the proposed test structure can be easily embedded by effectively taking advantage of the spare area between the two chips and accordingly area overhead is minimal.

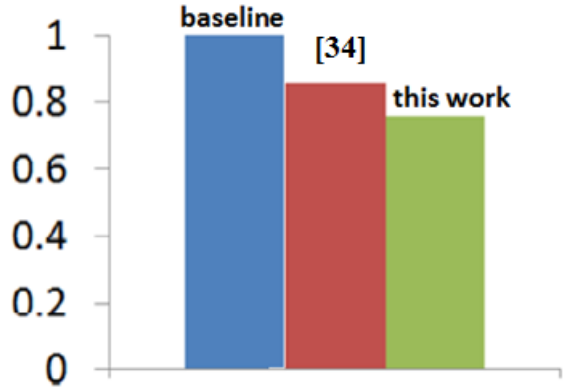


Figure 13. Simulated total training time (normalized to baseline slow case)

We measured the training time of NPU in various process corners with different interconnect delay. Figure 14 depicts the results for two chips: the slowest one measured and one that can potentially run faster. Note that we have normalized the total training time with regards to the baseline worst case. Based on our measurement, the performance of NPU varies significantly with changes in interconnect and transistor delay, which can be as large as 30% between the

best and the worst conditions. We estimated parasitic RC of chips in different corners by our on-chip monitor. We also used quiet-mode data to estimate the parasitic that would be reported by state-of-art test structure [34]. With conservative design methodology, we have to make the NPU work at the slowest clock frequency regardless of the actual interconnect delay, so the training time is always the same as the worst-case baseline. With state-of-art test structure [34], the gate capacitance and interconnect delay can be tracked but with large error (~27%). Therefore, the clock rate estimated from state-of-art test structure [34] is still suboptimal, even though it can achieve better performance than baseline (~15% improvement). With our test structure, we can track transistor and interconnect delay variation accurately, and clock frequency can be obtained. A performance improvement of 25% is obtained compared with baseline case.

# Chapter 6 Memory access in machine learning applications

## 6.1 Overview

In modern machine learning applications, SRAM plays a critical role in performance, as machine learning algorithms need intensive memory access. As an example, AlexNet [35], a deep learning neural network for image classification, is shown in Figure 15. AlexNet has five convolution layers, three max-pooling layers, two fully connected layers and one soft-max output layer.

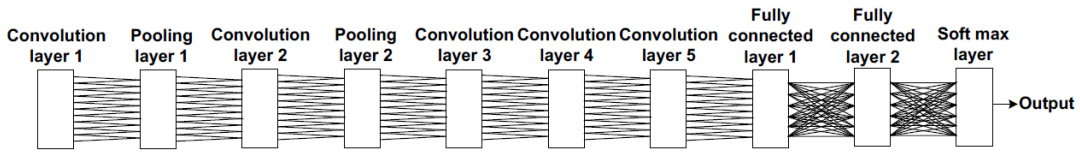


Figure 14. Structure of AlexNet

The most computationally intensive part of AlexNet convolution layer (Figure 15) and fully connected layer (Figure 16).

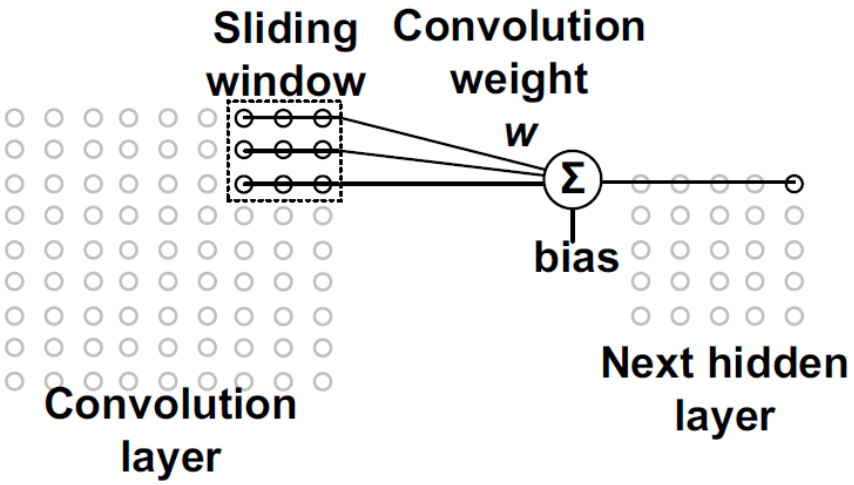


Figure 15. Convolution layer

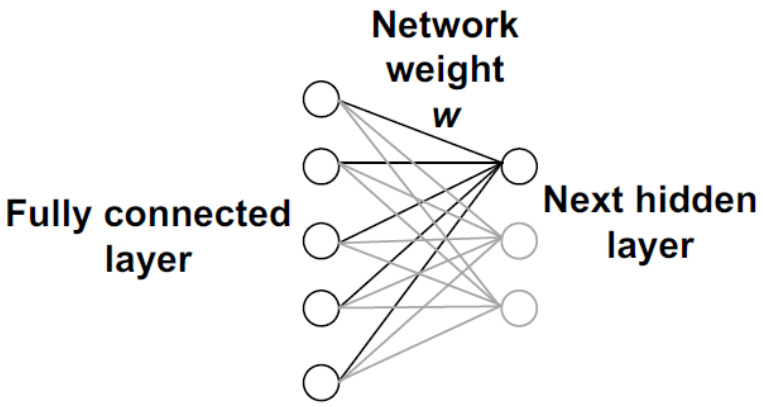


Figure 16. Connected layer

For each neuron in convolution layer, their input is the convolution of neurons in sliding window of the previous layer ($a$) with convolution weight ($w$) plus bias ($b$), which can be expressed as:

Equation 31. Neuron input

$$z_k = b_k + \sum_{j=0}^{N} w_{k,j} a_j$$

With their output as a non-linear function ($\sigma$) of input. For the convolution layer, convolution weight $w$ and bias $b$ are shared among neurons. The non-linear $\sigma$ can be sigmoid, tanh, rectified linear, etc. For a fully connected layer, the input of each neuron is the sum of the product of each neuron ($a$) in the previous layer with network weight ($w$), plus bias ($b$). This can be expressed as:

Equation 32. Neuron input for connected layer

$$z_k = b_k + \sum_{j=0}^{N} w_{k,j} a_j$$

Similar to the convolution layer, the output of each neuron in a fully connected layer is a non-linear function of input. For fully connected layer, network weight *w* and *b* are generally different for each neuron. The computation flow chart of the convolution layer and fully connected layer is shown in Figure 17. First, data of related neurons are fetched from memory, along with the corresponding weight and bias. After that, multiplication and accumulation (MAC) operation is performed with neuron, weight and bias as input. Then, the output of MAC is sent to non-linear function σ, which can be implemented with look-up table (LUT), and the final output of neuron is stored into memory again. For computation of each neuron, memory read and memory write operations are necessary. The data amount to read from memory for computation of each neuron can be large. For convolution layer, the sliding window size is usually between 5x5 [36] to 11x11 [35], and corresponding weight and bias also need to be fetched from memory. For fully connected layer, we need to read all neuron outputs (as many as 4096 in AlexNet) from previous layer along with corresponding network weights and bias. Therefore, memory access time can be critical for DL neuron network.

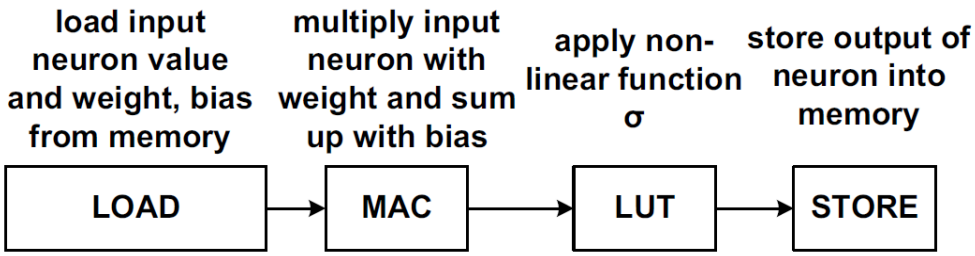


Figure 17. Flowchart of neuron computation in neural network

As another example, Vocabulary Forest (VF) [37] algorithm for video object recognition is shown in Figure 18. The VF algorithm has multiple Vocabulary Trees (VT), and it combines the output of VTs with different weights to calculate final match result. For each VT, input is key points of video frame extracted by scale invariant feature transform (SIFT). For each key point vector, its nearest neighbor of visual word is found by tree search algorithm (Figure 19)

from tree node table. The histogram of nearest neighbor visual words of the frame is obtained after all key points go through the VT. The final matched object of VF is calculated based on the histogram generated by all VTs. The most time-consuming part of VF algorithm is tree search for visual word and object match [38]. The computation flow chart of visual word search and object match can be abstracted in Figure 19. Again, each computation of visual word search and object match requires memory read to obtain tree node/object for comparison. Therefore, memory access time is also critical for VF algorithm.

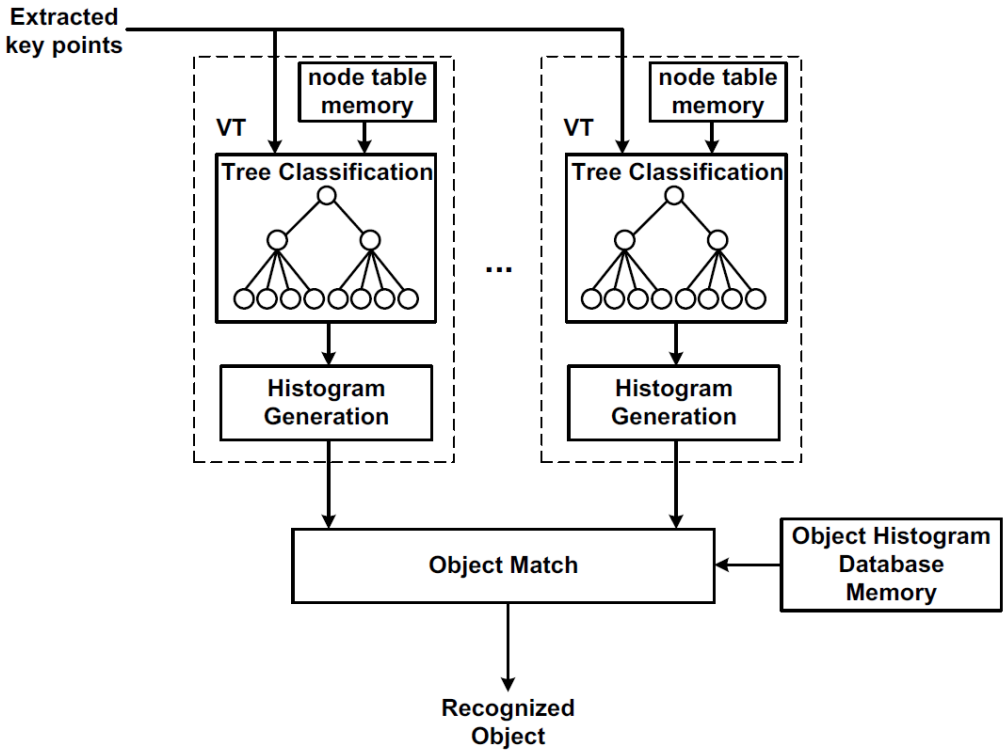


Figure 18. VF algorithm

Though SRAM is intensively accessed, different data in memory might be visited with various frequencies. For convolution layer in neural network, it shares the same bias and weight for all neurons. Therefore, the same weight and bias data need to be read from memory for neuron calculation in this layer. Consequently, the memory location where the data of weight and bias data are stored is frequently visited. On the other hand, the output of neuron is stored in memory once and then will not be touched again by this layer. Thus, the memory location

where neuron output is stored is less visited. For VF algorithm, the memory visit pattern depends on the object to be recognized in video frame. Part of VT nodes are less visited due to the fact that the tree nodes in this part of VT are less correlated to the target object [36], while the other part of VT nodes are more frequently visited since the tree nodes there are closer to the feature of the target object. As a result, the memory that stores the less- matched part of VT is not frequently visited, while the memory that stores the better-matched part of VT is more frequently visited.

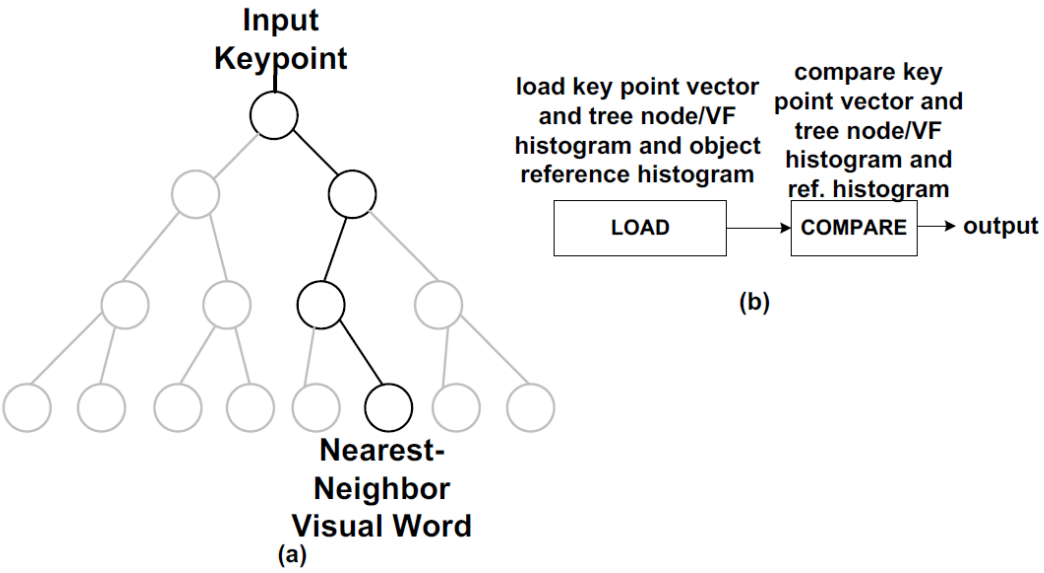


Figure 19. (a) Tree search algorithm of VT and (b) computation flow of visual word search/object histogram match

The difference in visit frequency of SRAM leads to variation in different parts of SRAM. Conventional design methodology requires every part of SRAM to pass the worst-case temperature variation condition, which means slower SRAM system clock. This methodology is too conservative, since only a small portion of SRAM may be in high temperature while most of SRAM is in nominal temperature and can run faster. The rationale behind conventional conservative design methodology is that in conventional design, there is no way to monitor temperature of SRAM in real time. In order to avoid functional failure, the design needs to be

pessimistic. To avoid the over-conservative design methodology, real-time temperature monitors for different SRAM parts need to be implemented, and the corresponding temperature compensation is required so that the performance of SRAM does not degrade even with significant temperature variation.

## 6.2 System architecture

In the proposed SRAM, fast access is achieved by folded SRAM structure and careful layout optimization. In addition, dual-loop process/temperature compensation is implemented to eliminate the effect of process and temperature to access time. Furthermore, the proposed SRAM uses one-port design instead of conventional dual-port and thereby saves routing resource. With optimized structure design, the proposed fast SRAM can run reliably at 1.8~2.2 GHz.

### 6.2.1 Folded structure

In the proposed SRAM, the control circuits (Pre-decoder, Row decoder, Dummy Bitcell, Global Control Unit) are placed in the spine of the memory bank as shown in Figure 20(a). The overall structure is a folded structure. Compared with conventional asymmetric structure (Figure 20(b)), the symmetric structure allows faster operation with lower energy consumption, since it has short word line length. The distributed wire RC is shown in Figure 21(a). From aspect of SRAM cell pre-charge time constant, the RC product for a specific length of the wire is fixed. With wider wire, unit-length resistance decreases linearly but unit-length capacitance increases linearly, and vice versa. Therefore, adjusting the width of word line can hardly influence the time constant. With a long wire, the speed is limited by the far-end SRAM cell. The rise time of near-end SRAM cells is short, while the rise time of far-end SRAM cells is long. Effectively, the timing margin of memory cell in near end is wasted (Figure 21(b)). Folded structure can avoid this waste and provide fast SRAM access.

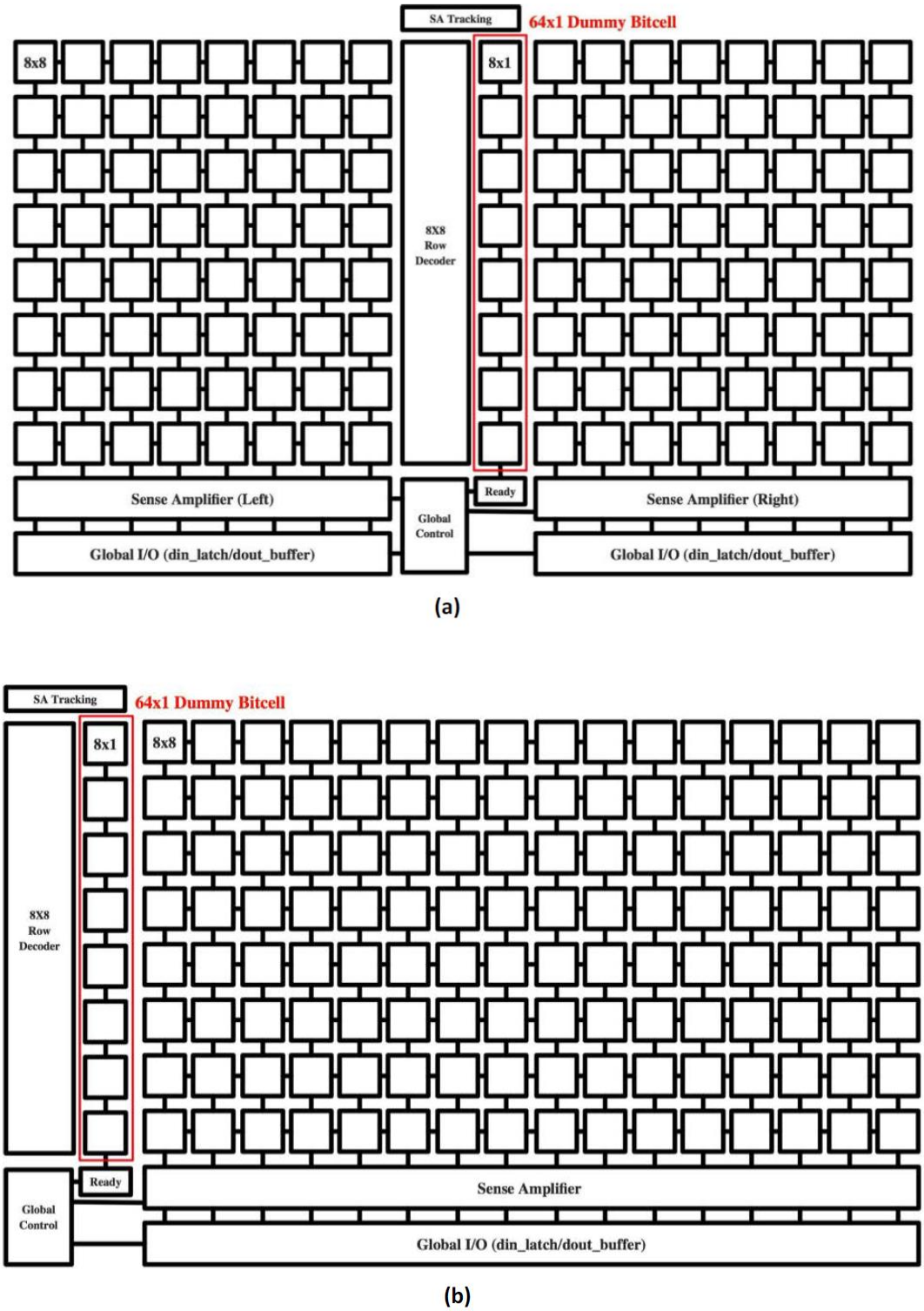


Figure 20. Folded (a) and conventional asymmetric (b) SRAM structure

In terms of energy consumption, the total system energy consumption after charging

loading capacitors to supply voltage Vdd is CVdd2. This value does not rely on wire resistance R. However, with the constraint of SRAM access time, a large RC constant requires the driver with higher Vdd. In other words, with small RC constant in the proposed folded structure, energy consumption can be lowered quadratically by decreasing Vdd, while satisfying access time constraints. In summary, the folded structure in our proposed SRAM system can provide a satisfying trade-off between speed and energy consumption.

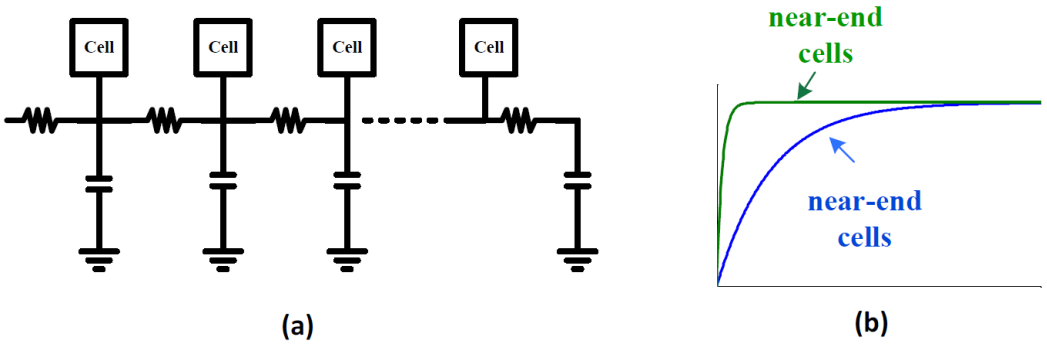


Figure 21. (a) Distributed RC in word line and (b) rise time of precharge.

### 6.2.2 Duel loop process/temperature compensation

In order to eliminate SRAM access time variation due to temperature and process drift, two key factors that are responsible for access time were identified by simulation. The first factor is RC load in word line, and the second is sensing amplifier (SA). Even though the proposed folded structure has short word line, the RC load variation is still significant under different process and temperature conditions. The RC load of word line can be much larger in the worst process and temperature conditions than in the typical case, and therefore SRAM speed degrades in the worst condition if no compensation is activated. In order to monitor word line RC load variation, a dummy 64x1 bit-cell column is implemented (Figure 20(a)). Every 8x1 dummy bit cell is responsible for monitoring the process and temperature variation of its row. When running the word line load compensation routine, global control unit (GCU) asserts a pulse through replica driver onto dummy bit-cell in each row. The timing counter in GCU then calculates the clock cycles of delay between pulse assertion and actual dummy word line pull-up/pull down. Based on the measured delay, GCU adjusts driving strength of word line driver for each row, and guarantees unchanged signal delay in SRAM word line even with process and temperature variation. The GCU replica driver-dummy-dummy bit-cell forms the first loop for process/temperature compensation (Figure 22).

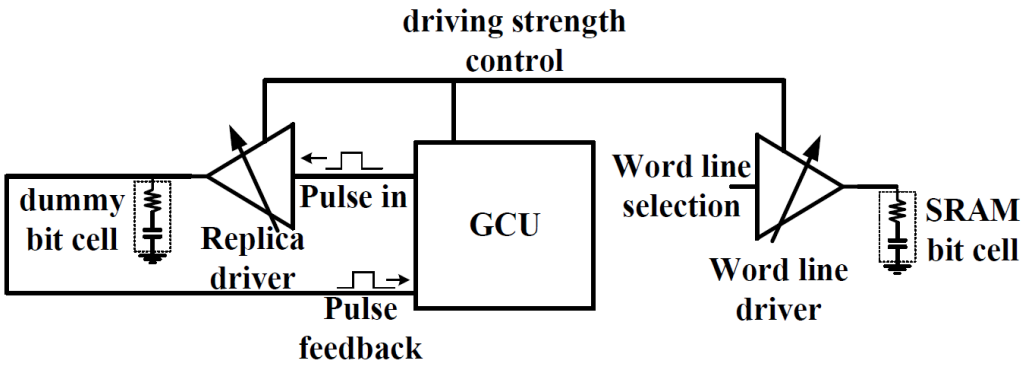


Figure 22. Proposed process/temperature compensation loop to compensate word line RC load variation

The second loop for process/temperature compensation eliminates delay variation of replica SA. The GCU sends a pulse to replica SA, and measures its output delay. The GCU adjusts the supply voltage of both replica and SRAM SA until measured delay meets timing requirement, as illustrated in Figure 23. The supply voltage adjustment is achieved by tuning the header resistance of SA. With tunable supply voltage based on the process and temperature conditions, SA can work with low power consumption when process and temperature are in typical condition. On the other hand, when process and temperature are in the worst condition, SA can still meet the constraint of delay by increasing supply voltage.

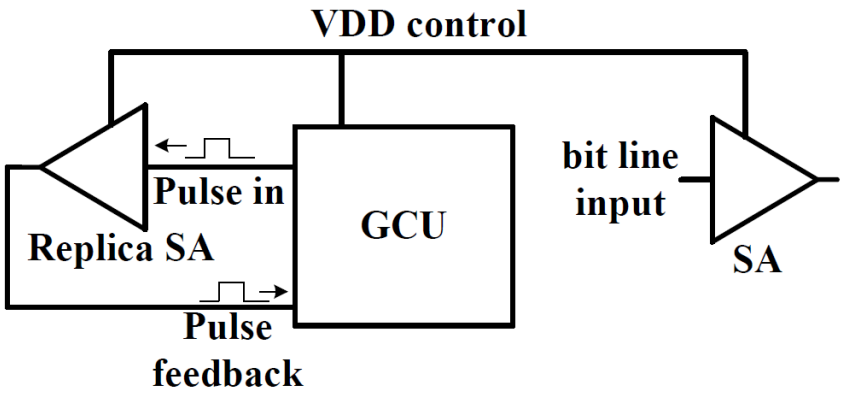


Figure 23. Proposed process/temperature compensation loop to compensate SA delay variation

Since only replica circuits and dummy cells are needed for loop to work, the proposed process/temperature compensation can compensate the process/temperature drift in real time without interrupting normal work of SRAM.

## 6.3 Circuit design details

### 6.3.1 Fast SRAM cells

The proposed SRAM cell structure is shown in Figure 24. A cross-coupled latch (M1-M4) acts as a one-bit storage and a pair of Pass Gates (M5, M6) is controlled by word line (WL). The bit line signals (BL and BLB) connect to all the cells in a column, working as a sense line in read operation and pull line in write operation.

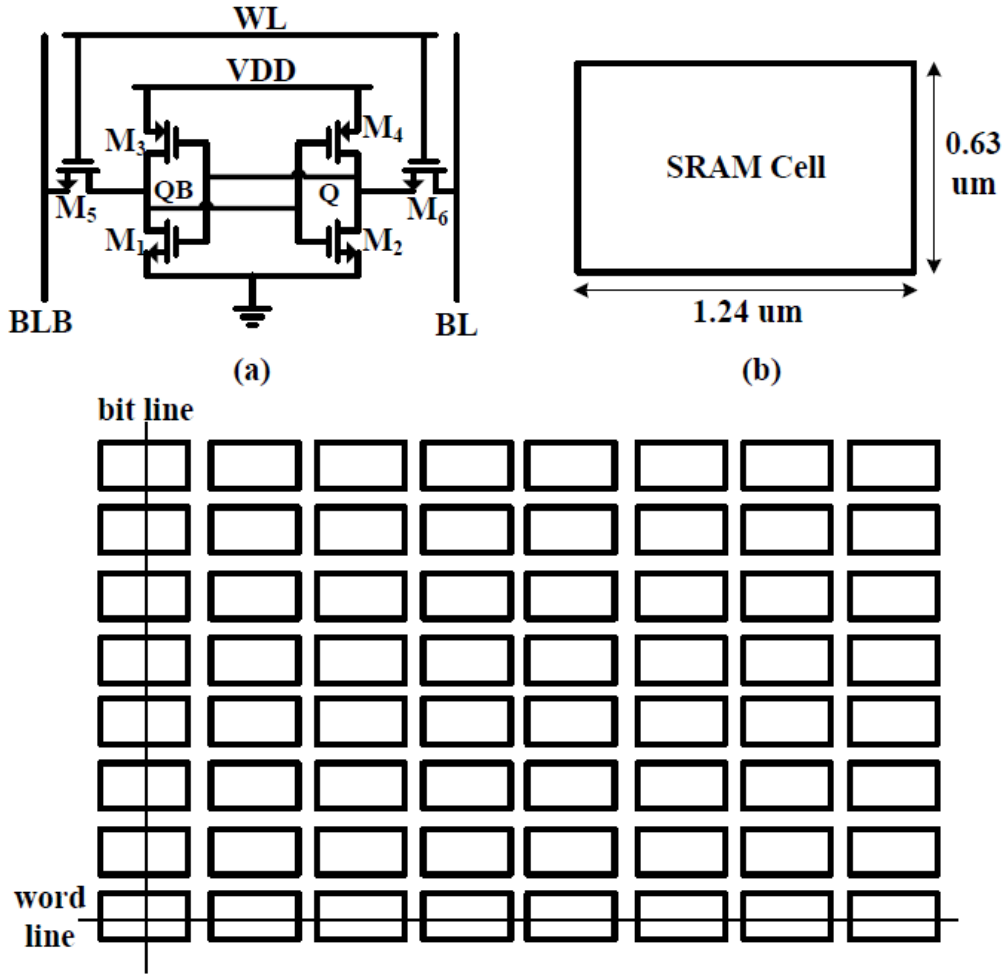


Figure 24. (a) SRAM cell schematic, (b) SRAM cell layout and (c) SRAM cell array layout

In order to optimize SRAM access speed, layout of SRAM cells is carefully optimized (Figure 24(b)). Unlike conventional SRAM design, the shape of SRAM cell is not square but rectangular. In the proposed SRAM cell array (Figure 24(c)), the word line is longer than bit

line. This reduces input load of SA, and the signal rise/fall time in bit line is improved. Therefore, SA can run at high speed without consuming too much power. On the other hand, word line is not exceedingly long as folded structure is used. Word line driver is also optimized to drive long word line.

### 6.3.2 Sensing amplifier

The proposed Sense Amplifier (SA) is shown in Figure 25. In the SA, M2-M5 is a cross couple latch, M6 – M8 work as pre-charging circuit, M9-M10 work as writing circuits, and M16-M24 work as reading circuits. The supply voltage in SA can influence speed and energy consumption of SA. With higher supply voltage, the SA has stronger driving capacity as MOSFETs have larger transconductance, and can therefore run faster. However, high speed comes at the cost of higher energy consumption for read and write operations, as SA need to charge capacitors to higher voltage [1]. In order to achieve balance between speed and energy consumption, the supply voltage of SA is designed as tunable with a header transistor (not drawn in Figure 25). When process and temperature are in typical corner, supply voltage is reduced to save energy. On the other hand, when process and temperature are in the worst corner, supply voltage is increased to meet speed requirements.

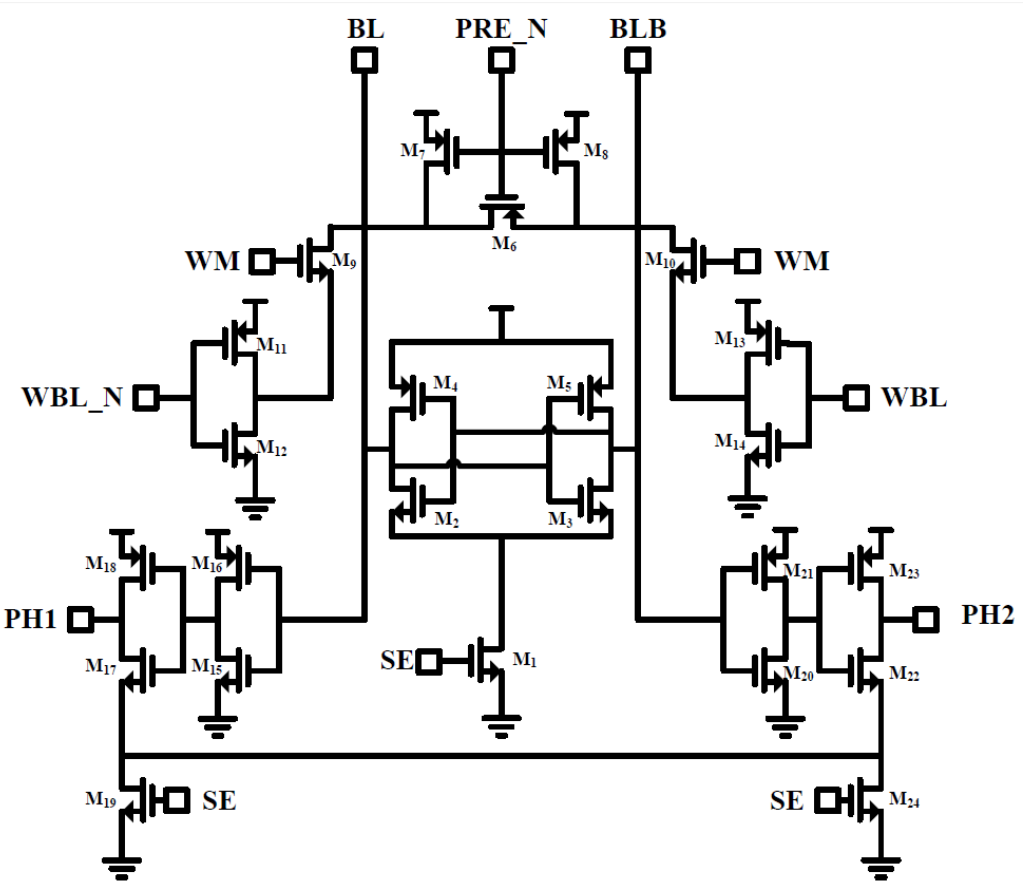


Figure 25. Proposed sensing amplifier

## 6.4 Implementation results

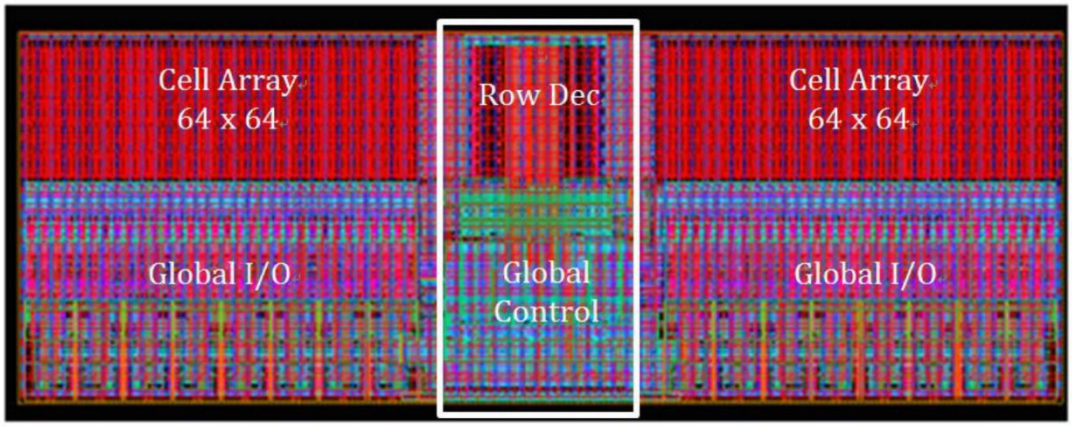


Figure 26. Proposed SRAM Layout view

The proposed 1-KB SRAM is implemented with Samsung 28nm LPP technology (Figure 27). Our design features a super-fast access time (2.2 GHz clock), which is 3X improvement over state-of-the-art design (Samsung HL 152). The cycle time is only 650 ps between two read cycles. The unit size of our SRAM is 121um x 43 um, which is 7% smaller than reference design. In addition, the pin port number is 12 for our design, which is more than 50% reduction compared with conventional dual-port design. The specification of the proposed SRAM is summarized in Table 5.

Table 5. Performance Summary

| **Design** | Samsung HL 152 | This Work |
|---|---|---|
| **Size (um x um)** | 127 X 44 | 121 X 43 |
| **# Pin Ports** | 25 | 12 |
| **Input Voltage** | 2 | 1 |
| **Speed** | 800MHz | 2.2GHz |
| **Power Consumption** | N/A | 120 pJ/access |

We evaluated temperature impact on the performance of our design by simulation. Our SRAM design is very robust against temperature and process variation, as shown in Figures 28 and 29. The x-axis in Figures 28 and 29 is the access time of different data channels (D0, D1, D2, D3 and column access CA). The access time variation in different temperature (0ºC~125ºC) and process (typical and worst) corners is less than 0.5 ns, which translates to less than 10% access time variation.

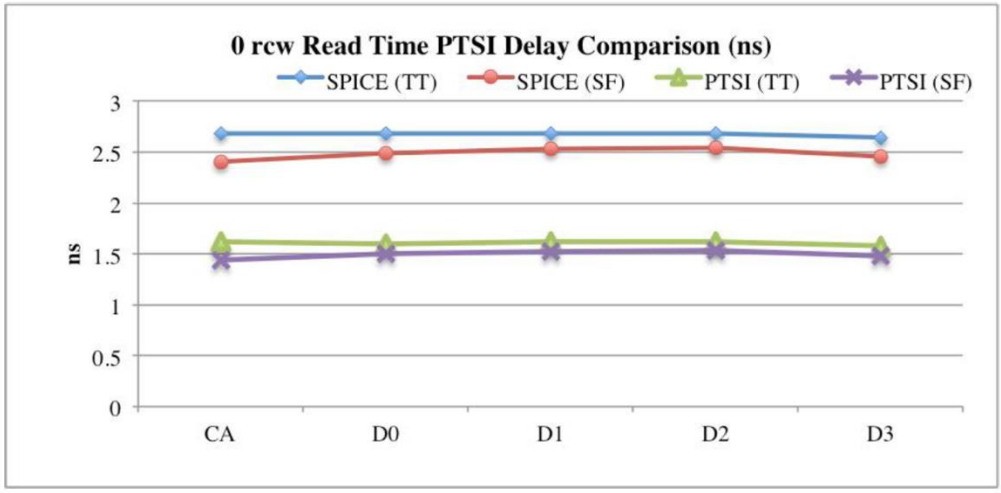


Figure 27. 0 rcw read time PTSI comparison result

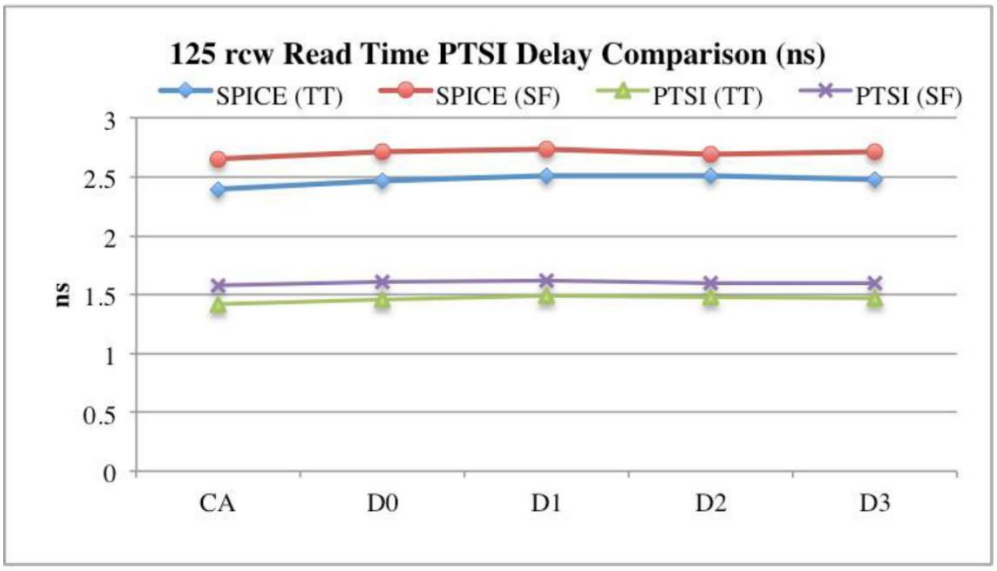


Figure 28. 125 rcw read time PTSI comparison result

The exemplary waveforms of read/write in 0ºC/125ºC are shown in Figure 30. The proposed SRAM can work reliably in extreme conditions without significant performance loss.

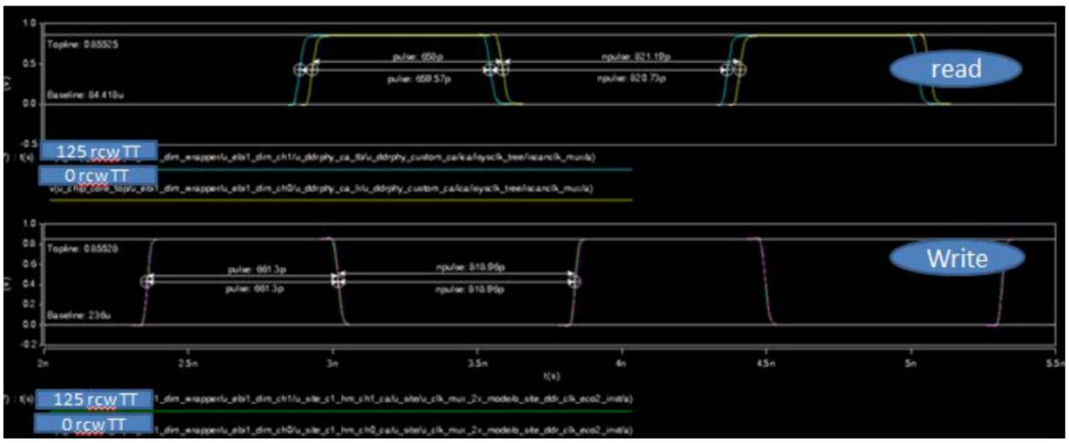


Figure 29. System architecture of the 20 Gb/s transceiver with the proposed SRAM integrated

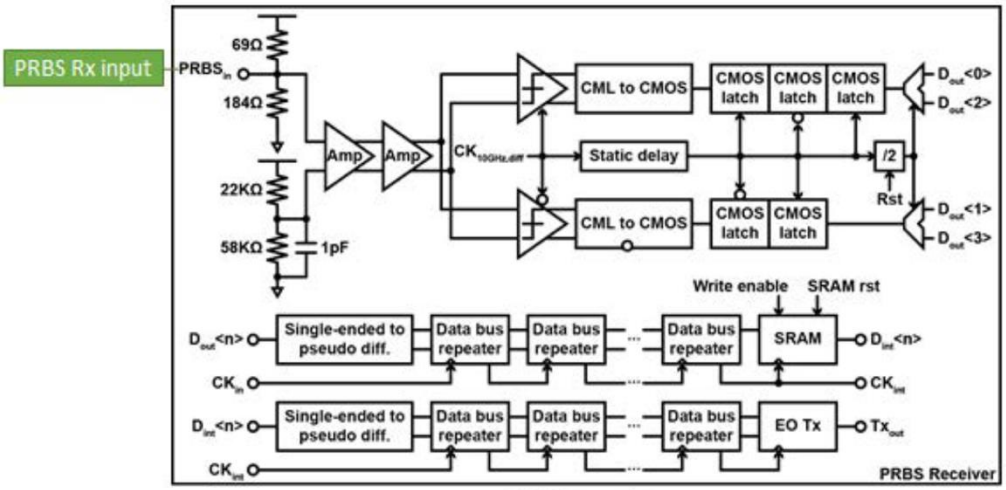


Figure 30. SERDES interface for transceiver/SRAM

The proposed SRAM is integrated as a building block of 20 Gb/s transceiver (Figure 30). The transceiver essentially sends serialized data from SRAM and writes received data back into SRAM and calculates bit-error-rate (BER). A SERDES is used to interface between fast SRAM and transceiver (Figure 31). The chip micrograph of transceiver with SRAM is shown in Figure 32, and the data eye diagram of the 20 Gb/s transceiver is shown in Figure 33. The clear eye diagram and low BER of transceiver verify that our fast SRAM work successfully at 2.2 GHz.

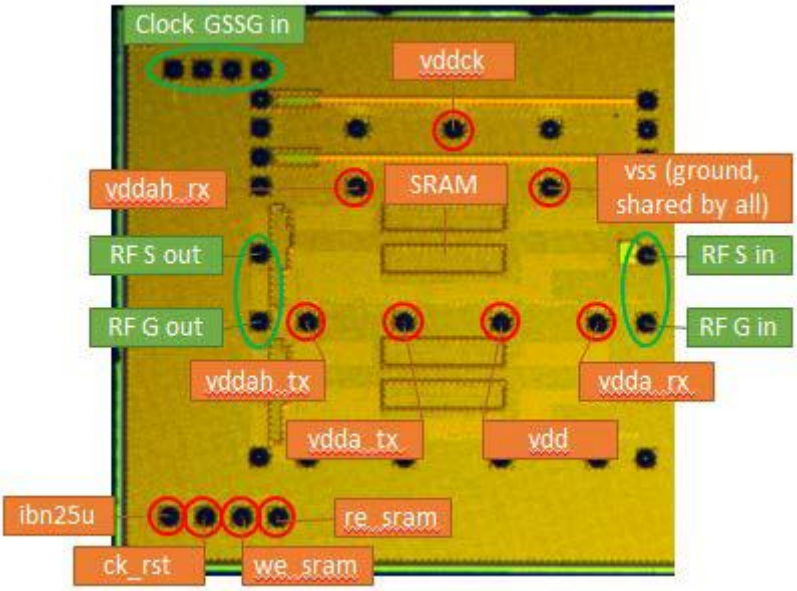


Figure 31. Micrograph of 20 Gb/s transceiver with the proposed SRAM integrated

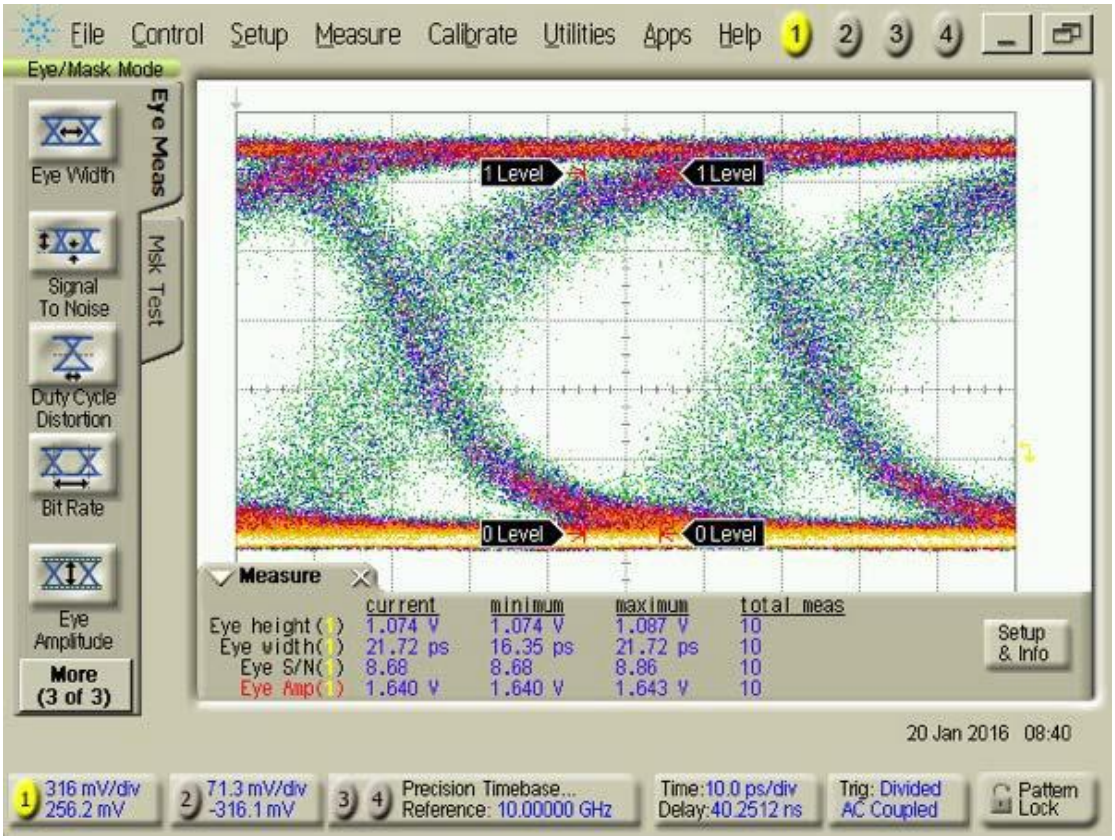


Figure 32. Measured eye diagram of 20 Gb/s transceiver with the proposed SRAM integrated

# Chapter 7 Conclusions and future works

In this thesis, we discuss the solutions proposed to improve the robustness of memory access and monitoring of the circuit uncertainty.

On our contribution on the monitoring of RC parasitic - since interconnect RC parasitic plays an important role to predict the overall chip performance for nanotechnology, enhanced test structures are implemented to measure the routing propagation delay and crosstalk for various interconnect configurations. A first-order empirical model is also developed to estimate RC parasitic. The test structures are quite simple and small. It can be easily inserted into real chip to monitor interconnect variation in volume production. Recently, we modified the proposed test structures to not only account for intra-layer (left/right routing) crosstalk but also for interlayer (top/bottom routing) crosstalk. Additional device parameters (i.e. source/drain capacitance) are extracted via minor test structure update. The modified test structures and empirical models are useful to estimate both device and interconnect parameters; it also help us to identify the serious interconnect process issues for clock tree design using TSMC 7nm FF process.

On the DFM analysis of analog circuit - guard ring and OD dummy fill effect on performance of MOSFET is studied in this thesis. The silicon measurement and simulation result difference is as high as 10% with guard ring and dummy fill so that they can be ignored. The preliminary guard ring and OD dummy fill models are proposed to improve simulation accuracy. For future development, more test keys are required to estimate the OD density gradient impacts and how to apply guard ring to protect the active circuit from the neighboring passive resistor and capacitor. Moreover, the measurement results have been taken into consideration for TSMC 7nm FF enhanced DFM guidelines for analog circuit yield enhancement.

Finally, to improve the memory access of machine learning hardware application, we

propose a new design of high-speed SRAM for machine learning purposes. With fast access time (cycle time: *650 ps*, access time: *350 ps*), low sensitivity to temperature variation and high reconfigurability (less than 10% performance difference between 125_rcw_tt vs 0_rcw_tt), the proposed SRAM is a better candidate for hardware machine learning system than the conventional SRAM. Compared with Samsung HL 152, our design has smaller size (*121X43 um2 vs 127X44 um2*) with half the number of pins ports (*12 vs 25*) and higher speed (*2.2GHz vs 800MHz).*

# Reference

[1] Alioto M: Ultra-low power VLSI circuit design demystified and explained: A tutorial. IEEE Transactions on Circuits and Systems I: Regular Papers 2012, 59(1):3-29.

[2] Rooseleer B, Cosemans S, Dehaene W: A 65 nm, 850 MHz, 256 kbit, 4.3 pJ/access, ultra low leakage power memory using dynamic cell stability and a dual swing data link. IEEE Journal of Solid-State Circuits 2012, 47(7):1784-1796.

[3] Rooseleer B, Dehaene W: A 40 nm, 454MHz 114 fJ/bit area-efficient SRAM memory with integrated charge pump. In: ESSCIRC (ESSCIRC), 2013 Proceedings of the: 2013: IEEE; 2013: 201-204.

[4] Tida UR, Zhuo C, Shi Y: Novel Through-Silicon-Via Inductor-Based On-Chip DC-DC Converter Designs in 3D ICs. ACM Journal on Emerging Technologies in Computing Systems (JETC) 2014, 11(2):16.

[5] Tida UR, Yang R, Zhuo C, Shi Y: On the efficacy of through-silicon-via inductors. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 2015, 23(7):1322-1334.

[6] Liu C, Su J, Shi Y: Temperature-aware clock tree synthesis considering spatiotemporal hot spot correlations. In: Computer Design, 2008 ICCD 2008 IEEE International Conference on: 2008: IEEE; 2008: 107-113.

[7] Nii K, Yabuuchi M, Tsukamoto Y, Ohbayashi S, Imaoka S, Makino H, Yamagami Y, Ishikura S, Terano T, Oashi T: A 45-nm bulk CMOS embedded SRAM with improved immunity against process and temperature variations. IEEE Journal of Solid-State Circuits 2008, 43(1):180-191.

[8] Liu CC, Wang Y-H, Li Y, Wong C-H, Chou TP, Chen Y-K, Chang M-CF: Invited-A 2.2 GHz SRAM with high temperature variation immunity for deep learning application under 28nm. In: Proceedings of the 53rd Annual Design Automation Conference: 2016: ACM; 2016: 2.

[9] Chen T, Du Z, Sun N, Wang J, Wu C, Chen Y, Temam O: Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In: ACM Sigplan Notices: 2014: ACM; 2014: 269-284.

[10] Bojnordi MN, Ipek E: Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning. In: 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA): 2016: IEEE; 2016: 1-13.

[11] Neshatpour K, Malik M, Homayoun H: Accelerating machine learning kernel in hadoop using fpgas. In: Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on: 2015: IEEE; 2015: 1151-1154.

[12] Kaul H, Anders MA, Mathew SK, Chen G, Satpathy SK, Hsu SK, Agarwal A, Krishnamurthy RK: 14.4 A 21.5 M-query-vectors/s 3.37 nJ/vector reconfigurable k-nearest-neighbor accelerator with adaptive precision in 14nm tri-gate CMOS. In: 2016 IEEE International Solid-State Circuits Conference (ISSCC): 2016: IEEE; 2016: 260-261.

[13] Park S, Bong K, Shin D, Lee J, Choi S, Yoo H-J: 4.6 A1. 93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications. In: 2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers: 2015: IEEE; 2015: 1-3.

[14] Gupta PK: Xeon+ fpga platform for the data center. In: Fourth Workshop on the Intersections of Computer Architecture and Reconfigurable Logic: 2015; 2015.

[15] Banerjee K, Im S, Srivastava N: Interconnect modeling and analysis in the nanometer era: Cu and beyond. In: 22nd advanced metallization conference: 2005: Citeseer; 2005.

[16] Aikawa H, Morifuji E, Sanuki T, Sawada T, Kyoh S, Sakata A, Ohta M, Yoshimura H, Nakayama T, Iwai M: Variability aware modeling and characterization in standard cell in 45 nm CMOS with stress enhancement technique. In: 2008 Symposium on VLSI Technology: 2008: IEEE; 2008: 90-91.

[17] Tsuno H, Anzai K, Matsumura M, Minami S, Honjo A, Koike H, Hiura Y, Takeo A, Fu W, Fukuzaki Y: Advanced analysis and modeling of MOSFET characteristic fluctuation caused by layout variation. In: 2007 IEEE Symposium on VLSI Technology: 2007: IEEE; 2007: 204-205.

[18] Hook TB, Brown J, Cottrell P, Adler E, Hoyniak D, Johnson J, Mann R: Lateral ion implant straggle and mask proximity effect. IEEE Transactions on Electron Devices 2003, 50(9):1946-1951.

[19] Wang L-N, Xu N, Toh S-O, Neureuther AR, Liu T-JK, Nikolic B: Parameter-specific ring oscillator for process monitoring at the 45 nm node. In: Custom Integrated Circuits Conference (CICC), 2010 IEEE: 2010: IEEE; 2010: 1-4.

[20] Wang LT-N, Pang L-T, Neureuther AR, Nikolić B: Parameter-specific electronic measurement and analysis of sources of variation using ring oscillators. In: SPIE Advanced Lithography: 2009: International Society for Optics and Photonics; 2009: 72750L-72750L-72710.

[21] Schroder DK: Semiconductor material and device characterization: John Wiley & Sons; 2006.

[22] Stavitski N, Klootwijk JH, van Zeijl HW, Kovalgin AY, Wolters RA: Cross-bridge Kelvin resistor structures for reliable measurement of low contact resistances and contact interface characterization. IEEE Transactions on Semiconductor Manufacturing 2009, 22(1):146-152.

[23] Bhaskaran M, Sriram S, Holland AS: Accurate Estimation of Low Values of Specific Contact Resistivity. IEEE Electron Device Letters 2008, 29(3):259-261.

[24] Bhushan M, Gattiker A, Ketchen MB, Das KK: Ring oscillators for CMOS process tuning and variability control. IEEE Transactions on Semiconductor Manufacturing 2006, 19(1):10-18.

[25] Liu C, Law OMK, Lu J-Y, Chen P-H, Duan Z: Uncertainty aware interconnect design to improve circuit performance and/or yield. In.: Google Patents; 2015.

[26] Tsai M-T, Huang S-Y, Tsai K-H, Cheng W-T: Monitoring the delay of long interconnects via distributed TDC. In: Test Conference (ITC), 2015 IEEE International: 2015: IEEE; 2015: 1-9.

[27] Woo A, Eberhart H, Li Y, Ito A: Mismatch in high-K metal gate process analog design. In: 2014 IEEE International Electron Devices Meeting: 2014: IEEE; 2014: 18.12. 11-18.12. 14.

[28] Razavi B, 罗扎: Design of analog CMOS integrated circuits: 清华大学出版社有限公司; 2001.

[29] Hastings RA: The art of analog layout: Prentice Hall; 2006.

[30] Kuhn K, Kenyon C, Kornfeld A, Liu M, Maheshwari A, Shih W-k, Sivakumar S, Taylor G, VanDerVoorn P, Zawadzki K: Managing Process Variation in Intel's 45nm CMOS Technology. Intel Technology Journal 2008, 12(2).

[31] Webb C: 45nm Design for Manufacturing. Intel Technology Journal 2008, 12(2).

[32] Xue J, Ye Z, Deng Y, Wang H, Yang L, Yu Z: Layout-dependent STI stress analysis and stress-aware RF/analog circuit design optimization. In: Proceedings of the 2009 International Conference on Computer-Aided Design: 2009: ACM; 2009: 521-528.

[33] Wei Y, Hu J, Liu F, Sapatnekar SS: Physical design techniques for optimizing RTA-induced variations. In: Proceedings of the 2010 Asia and South Pacific Design Automation Conference: 2010: IEEE Press; 2010: 745-750.

[34] Fujimoto S, Islam AM, Matsumoto T, Onodera H: Inhomogeneous ring oscillator for within-die variability and RTN characterization. IEEE Transactions on Semiconductor Manufacturing 2013, 26(3):296-305.

[35] Krizhevsky A, Sutskever I, Hinton GE: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems: 2012; 2012: 1097-1105.

[36] LeCun Y, Bottou L, Bengio Y, Haffner P: Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998, 86(11):2278-2324.

[37] Lee KJ, Kim G, Park J, Yoo H-J: A vocabulary forest object matching processor with 2.07 M-vector/s throughput and 13.3 nJ/vector per-vector energy for full-HD 60 fps video object recognition. IEEE Journal of Solid-State Circuits 2015, 50(4):1059-1069.

[38] Chen T-W, Su Y-C, Huang K-Y, Tsai Y-M, Chien S-Y, Chen L-G: Visual vocabulary processor based on binary tree architecture for real-time object recognition in full-HD

resolution. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 2012, 20(12):2329-2332.

[39] A. Tang et al., "A 65nm CMOS 140 GHz 27.3 dBm EIRP transmit array with membrane antenna for highly scalable multi-chip phase arrays," IEEE International Microwave Symposium 2014.

[40] A. Tang et al., "A 200 GHz 16-pixel focal plane array imager using CMOS super regenerative receivers with quench synchronization," IEEE International Microwave Symposium 2012.

[41] H. Wu et al., "A Current-Mode mm-Wave direct-conversion receiver with 7.5GHz Bandwidth, 3.8dB minimum noise-figure and +1dBm P1dB, out linearity for high data rate communications," 2013 IEEE Radio Frequency Integrated Circuits Symposium (RFIC), Seattle, WA, 2013, pp. 89-92.

[42] Z. Z. Chen et al., "A wide-band 65nm CMOS 28–34 GHz synthesizer module enabling low power heterodyne spectrometers for planetary exploration," 2015 IEEE International Microwave Symposium, AZ, 2015, pp. 1-3.

[43] Y. Zhao et al., "A 0.56 THz Phase-Locked Frequency Synthesizer in 65 nm CMOS Technology," in IEEE Journal of Solid-State Circuits, vol. 51, no. 12, pp. 3005-3019, Dec. 2016.

[44] R. Al Hadi, Y. Zhao, Y. Li, Y. Du, and M.-C. F. Chang, "Retroactive terahertz displacement sensor in a standard 65nm CMOS technology", in Proc. OSA Conf. Lasers and Electro-Optics (CLEO), San Jose, CA, Jun. 2016, pp. 1-3.

[45] R. A. Hadi et al., "A spectral profiling method of mm-wave and terahertz radiation sources," 2016 IEEE International Microwave Symposium, CA, 2016, pp. 1-3.

[46] H. Wu, N. Y. Wang, Y. Du and M. C. F. Chang, "A Blocker-Tolerant Current Mode 60-GHz Receiver With 7.5-GHz Bandwidth and 3.8-dB Minimum NF in 65-nm CMOS," in IEEE Transactions on Microwave Theory and Techniques, vol. 63, no. 3, pp. 1053-1062, March 2015.

[47] Y. Li et al., "A multi-band low-noise transmitter with digital carrier leakage suppression and linearity enhancement," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 5, pp. 1209-1219, May 2013.

[48] Z. Z. Chen et al., "Digital PLL for phase noise cancellation in ring oscillator-based I/Q receivers," 2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits), June 2016.

[49] A. Tang et al., "A 95 GHz centimeter scale precision confined pathway system-on-chip navigation processor for autonomous vehicles in 65nm CMOS," IEEE International Microwave Symposium 2015.

[50] A. Tang et al., "CMOS (Sub)-mm-Wave System-on-Chip for exploration of deep space and outer planetary systems," IEEE International Microwave Symposium 2015.

[51] Lv Jingjing Du Li(School of Information Science and Engineering of Southeast University,Nanjing 210096,China);Vehicular Collision Avoiding System Based on Two Ultrasonic Receivers[J];Value Engineering;2010-22

[52] L. Du, Y. Zhang, C. C. Liu, A. Tang, F. Hsiao and M. C. F. Chang, "A 2.3-mW 11-cm Range Bootstrapped and Correlated-Double-Sampling Three-Dimensional Touch Sensing Circuit for Mobile Devices," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 1, pp. 96-100, Jan. 2017.

[53] L. Du, Y. Zhang, F. Hsiao, A. Tang, Y. Zhao, Y. Li, J. Chen, L. Huang, M.-C. F. Chang, "A 2.3mW 11cm Range Bootstrapped and Correlated Double Sampling (BCDS) 3D

Touch Sensor for Mobile Devices," IEEE International Solid-State Circuits Conference, pp. 122-123, Feb. 22-26, 2015.

[54] L. Du et al., "Invited — Airtouch: A novel single layer 3D touch sensing system for human/mobile devices interactions," 2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, 2016.

[55] W. H. Cho et al., "A 5.4-mW 4-Gb/s 5-band QPSK transceiver for frequency-division multiplexing memory interface," 2015 IEEE Custom Integrated Circuits Conference (CICC), San Jose, CA, 2015, pp. 1-4.

[56] W. H. Cho, et al., “A 38mW 40Gb/s 4-lane tri-band PAM-4/16-QAM transceiver in 28nm CMOS for high-speed Memory interface,” IEEE ISSCC Dig. Tech. Papers, pp. 184-185, Feb. 2016.

[57] Y. Du et al., "A 16Gb/s 14.7mW tri-band cognitive serial link transmitter with forwarded clock to enable PAM-16 / 256-QAM and channel response detection in 28 nm CMOS," 2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits), Honolulu, HI, 2016, pp. 1-2.

[58] Y. Du et al., "A 16-Gb/s 14.7-mW Tri-Band Cognitive Serial Link Transmitter With Forwarded Clock to Enable PAM-16/256-QAM and Channel Response Detection," in IEEE Journal of Solid-State Circuits , vol.PP, no.99, pp.1-12.

[59] B. Hu et al., "A Capacitor-DAC-Based Technique For Pre-Emphasis-Enabled Multi-Level Transmitters," in IEEE Transactions on Circuits and Systems II: Express Briefs , vol.PP, no.99, pp.1-1

[60] Du, Yuan. (2016). Cognitive Serial Interface with Multi-Band Signaling and Channel Learning Mechanism. UCLA: Electrical Engineering 0303. Retrieved from: http://escholarship.org/uc/item/8vs373c5

[61] Y. Li et al., "Transconductance enhancement method for operational transconductance amplifiers," Electronics Letters, vol. 46, no. 19, pp1321-1323, 2010.

[62] Y. Li et al., "Analysis and implementation of an improved recycling folded cascode amplifier," Journal of Semiconductors, vol. 33, no. 2, 2012.

[63] Y. Li et al., "A subthreshold MOSFET bandgap reference with ultra-low power supply voltage," IEEE 9th International Conference on ASIC (ASICON), Shanghai, 2011.

[64] Y. Li et al., "A 0.6 ppm/C current-mode bandgap with second-order temperature compensation," IEEE 9th International Conference on ASIC (ASICON), Shanghai, 2011.

[65] Y. Li et al., "Carrier synchronisation for multiband RF interconnect (MRFI) to facilitate chip-to-chip wireline communication," in Electronics Letters, vol. 52, no. 7, pp. 535-537, 2016.

[66] L. Du, "An Overview of Mobile Capacitive Touch Technologies Trends",2016, arXiv:1612.08227 [cs.ET]

[67] C. C. Liu, L. Du et al., "A single layer 3D Touch Sensing System for Mobile Devices Application," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol.PP, no.99, pp.1-12

[68] Yanghyo Kim et al., “Baseband Equivalent Impulse Response Analysis of Waveguides”, in IEEE Transactions on Microwave Theory and Techniques, vol. pp, no.99, pp. 1-12, 2016.

[69] Yanghyo Kim et al., “Impulse Response Analysis of Carrier-Modulated Multiband RF-Interconnect (MRFI)”, in IEEE Transactions on Microwave Theory and Techniques, vol. pp, no.99, pp. 1-12, 2016.

[70] Yanghyo Kim et al., “A 125GHz Transceiver in 65nm CMOS Assembled with FR4 PCB Antenna for Contactless Wave-Connectors", 2017 IEEE International Microwave Symposium, HA, 2017, pp. 1-3.

[71] A. Tang et al.," Chirp-Partition based Pre-Distortion for Reduced Carrier Leakage in Circulator-based Wide-band FMCW Radar Systems ",2017 IEEE International Microwave Symposium, HA, 2017, pp. 1-3.

[72] Tong Zhang et al., "A Simple System for Measuring Antenna Radiation Patterns in the Wi-Fi Band," in IEEE Antennas and Propagation Magazine, vol. 55, no. 1, pp. 191-202, Feb. 2013.

[73] CC Liu, O Lau, JY Du, "Complete DFM Model for High-Performance Computing SoCs with Guard Ring and Dummy Fill Effect", arXiv preprint arXiv:1701.00460

[74] CC Liu, O Law, F Li, "An Accurate Interconnect Test Structure for Parasitic Validation in On-Chip Machine Learning Accelerators", arXiv preprint arXiv:1701.03181

[75] CC Liu, YH Wang, Y Li, CH Wong, TP Chou, YK Chen, MCF Chang, "A 2.2 GHz SRAM with high temperature variation immunity for deep learning application under 28nm", Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE, 1-6

[76] CC Liu, J Su, Y Shi, "Temperature-aware clock tree synthesis considering spatiotemporal hot spot correlations", Computer Design, 2008. ICCD 2008. IEEE International Conference on, 107-113

[77] CC Liu, JY Lu, S Xie, "Mitigating electromigration, in-rush current effects, IR-voltage drop, and jitter through metal line and via matrix insertion", US Patent 9,496,174

[78] CC Liu, S Xie, "Time-variant temperature-based 2-D and 3-D wire routing", US Patent 9,298,874