Chapter 5
Variations in Device Characteristics

Hidetoshi Onodera, Yukiya Miura, Yasuo Sato, Seiji Kajihara, Toshinori Sato, Ken Yano, Yuji Kunitake and Koji Nii

Abstract Ever increasing variability in device characteristics is a major threat to the dependability, since it could give rise to faults and failures in VLSI circuits and systems. The variability arises from the variation in device parameters, such as geometry and doping densities, that is inherently associated with the technology scaling. This chapter deals with the variability of scaled devices and countermeasures to enhance dependability both at the device and circuit levels. First, in Sect. 5.1, variations in transistor characteristics are overviewed with measured variability from 0.35 µm down to 40 nm technologies. The rapid increase in within-die random variations is clearly shown. Possible scaling scenarios, which are device-level strategies to reduce variability, are explained. In the following sections, we discuss countermeasure techniques at the circuit level. In Sect. 5.2, on-chip monitor circuits for variability measurement and performance compensation by localized body biasing are proposed and verified by silicon measurements. In Sects. 5.3 and 5.4, two techniques for predicting and preventing timing faults during runtime are introduced. The first technique in Sect. 5.3 relies on accurate delay-time measurement by an on-chip monitor circuit. Timing margins reduced by

H. Onodera (✉)
Kyoto University, Kyoto, Japan
e-mail: onodera@vlsi.kuee.kyoto-u.ac.jp

Y. Miura
Tokyo Metropolitan University, Hino, Japan

Y. Sato · S. Kajihara
Kyushu Institute of Technology, Iizuka, Japan

T. Sato
Fukuoka University, Fukuoka, Japan

K. Yano
Tokyo Institute of Technology, Tokyo, Japan

Y. Kunitake
Panasonic Corporation, Kadoma, Japan

K. Nii
Renesas Electronics Corporation, Tokyo, Japan

S. Asai (ed.), *VLSI Design and Test for Systems Dependability*,
https://doi.org/10.1007/978-4-431-56594-9_5

aging effects such as negative-bias-temperature instability (NBTI) can be evaluated and compensated. The second technique in Sect. 5.4 proposes a warning flip-flop that can predict possible timing errors before they actually happen, thus enables dependable operation throughout the whole life cycle of the circuit. Finally in Sect. 5.5, variability-aware circuit architectures are discussed for Static Random Access Memories (SRAMs). The proposed SRAM achieves expanded operating margins by fine-grain assist bias control at low supply voltages.

Keywords Device variation • Process variation • On-chip monitor
Variation-aware design • Timing error prediction and compensation

5.1 Overview of Device Variations

Hidetoshi Onodera, Kyoto University

5.1.1 Device Variation and Overview of This Chapter

With the device dimensions in the nanometer regime, variability in device performance becomes a crucial problem in LSI design. The variability comes from the physical-level fluctuations in device structures, and appears as the fluctuations in device characteristics such as drain currents and threshold voltages, and leads to the variations in circuit-level performances such as delay and power dissipation. These "faults" may cause "errors" which eventually result in malfunctions ("failures") of LSI circuits and systems.

The variability, however, is not a new problem and it has been always an issue in circuit design. In the past, the variability mainly came from imperfect control of fabrication processes. Device performance varied from a lot to lot and from a wafer to wafer, while the variation within a die was relatively small. We can say that the variability had a "global" nature and a local fluctuation within a die could be neglected in many cases except for certain analog designs. Although the amount of global variation could be large, the global nature allows us to evaluate the effect of variation by the worst-case analysis where all the devices are assumed to have the performance of the same extreme corner. In this way, the global variability has been managed mainly considering the performance at all the worst-case corners of device performances. On the other hand, the variability in the present and the future have different statistical characteristics. As device dimensions have been approaching atomic scales, intrinsic atomistic variations such as line edge roughness and discrete random dopant fluctuations become prominent [1, 2]. Those atomistic variations are random in nature and result in a random within-die variation of device performances. The random variation cannot be handled by the worst-case analysis since the possibility of all the devices being at the same worst corner becomes extremely small. The worst-case design becomes unrealistically pessimistic and results in the reduced advantage of scaling. It is, therefore, important to establish a new design methodology that considers the statistical nature of the variation.

This section, focusing on MOS transistors in scaled technologies, gives an overview of device variations and possible solutions at the device level. In Sect. 5.1.2, we classify the variations from a standpoint of spatial distributions. In Sect. 5.1.3, we explain the sources of variations. In Sect. 5.1.4, we show examples of measured variations from 0.35 μm to 40 nm technologies which indicate the statistical nature of the variation is changing from the global to local, in other words, from die-to-die to within-die. In Sect. 5.1.5, we discuss the variability trend and possible scaling scenarios for the future, which provides a countermeasure against variability at the device level. Section 5.1.6 summarizes this section.

Following sections cover a variety of circuit-level proposals that cope with device variations. Section 5.2 explains a method for monitoring variations and a countermeasure for compensating the variations. Section 5.3 proposes an on-chip monitor circuit for accurate delay-time measurement and Sect. 5.4 introduces a warning flip-flop that can predict timing errors due to delay variation in a critical path. Section 5.4 introduces a circuit design technique and a clocking scheme that can overcome timing faults due to variations. The last section of this chapter, Sect. 5.5, proposes design techniques that enhance the operating margins of Static Random Access Memory (SRAM) under device variations.

5.1.2 Classification of Variation

Components of performance variations can be classified into “global” one and “local” one from a standpoint of spatial distributions. The global component gives a uniform variation in the same direction to all the transistors on a die, and is called a D2D (Die-to-Die) variation. Lot-to-lot and wafer-to-wafer variations correspond to D2D variations. The local component arises in each individual transistor and is called a WID (Within-die) variation. From a standpoint of statistical nature, WID variations can be further classified into those that gradually fluctuate over a chip and those that randomly fluctuate [2].

5.1.3 Sources of Variation

The width of the gate, the oxide thickness, and the dopant density of a transistor have direct effects on the transistor performance. Fluctuations in the dimension of device structures mainly appear as D2D variations. It is important that those fluctuations should be maintained within the worst corners so that enough yields are achieved. In a scaled technology, besides those historical sources of variations, atomistic-level variations in an individual transistor such as discrete random dopant fluctuations (RDF: Random Dopant Fluctuation) and fluctuations in a sidewall of gate material (LER: Line Edge Roughness) have a strong impact on performance variations. Due to the statistical nature of those sources, they appear as random components in WID variations. Stress variations in strained Si and STI (Shallow Trench Isolation) processes, dopant density fluctuations near well boundaries due to dopant scattering during well-forming processes, and local temperature variations during rapid thermal annealing come into play as the sources of variation. Those sources contribute to location-specific components in WID variations. Aggressive scaling and increasing technology complexity lead to an explosion in the magnitude of variability while also introducing new sources of variations.

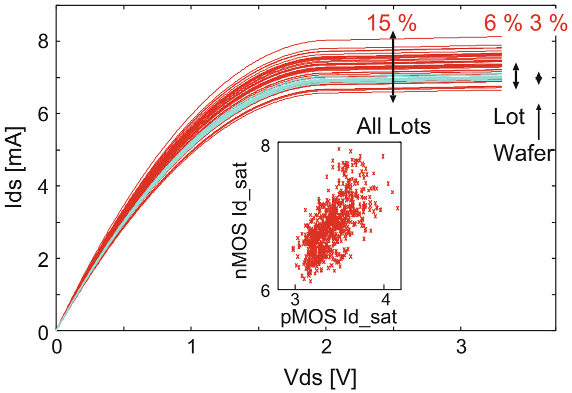
5.1.4 Observation of Variation

We examine variability trend from measured characteristics of five different fabrication technologies: 0.35 μm, 180 nm, 90 nm, 65 nm, and 40 nm. We can see a growing trend of WID variability as the scaling advances.

5.1.4.1 Evaluation of Variation

As an example of variations in the past technology, we show variations of transistor characteristics in a 0.35 μm process. The drain saturation current ($I_{\mathrm{d\_sat}}$) and the threshold voltage (V_{th}) of 16 PCM (Process Control Module) transistors distributed over a wafer have been measured for 58 lots with 797 wafers. Figure 5.1 shows the distribution of drain current characteristics reconstructed from the measured $I_{\mathrm{d\_sat}}$ and V_{th}. If we superimpose characteristics for all lots, the 3σ width of $I_{\mathrm{d\_sat}}$ variations becomes about 15% of the mean $I_{\mathrm{d\_sat}}$ value. If we build up for a single lot only, the average of the 3σ width is reduced to about 6%. If we consider 16

Fig. 5.1 Drain current variations in a 0.35 μm process



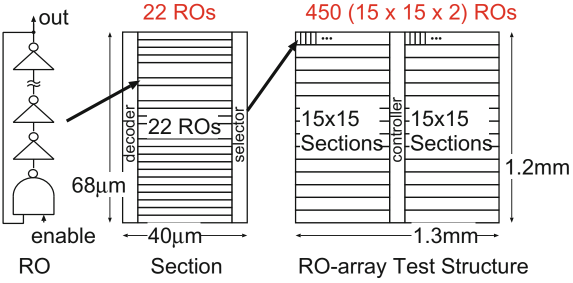


Fig. 5.2 RO (Ring Oscillator)-array test structure for variability characterization

transistors over a single wafer, the average of the 3σ width becomes 3%. We do not have data for estimating within-die variation. However, it is expected that the D2D component dominates over the WID component. A scatter plot inside Fig. 5.1 shows a distribution of nMOS $I_{\mathrm{d\_sat}}$ and pMOS $I_{\mathrm{d\_sat}}$. In this process generation, we can safely rely on a corner-based design method.

We next show variations of oscillation frequencies that are measured from an array of ROs (Ring Oscillators) fabricated in four technology generations of 180, 90, 65, and 40 nm. An example of the RO array circuit for variability characterization is shown in Fig. 5.2. This circuit is fabricated in a 90 nm technology. A variety of ROs is assembled in a block called “Section.” The circuit in Fig. 5.2 includes 22 types of ROs in a Section. The Section is then arranged in two sets of a 15-by-15 array, resulting in 450 ROs for each circuit configuration. The size of the test structure is 1.2 mm by 1.3 mm. If there is no variation in device characteristics, all the ROs with the same circuit configuration should have the same oscillation frequency. However, due to D2D and WID variations, the oscillation frequency of each RO varies. With this test structure, we can estimate the variability of D2D and WID components in a form of the oscillation frequency. Figure 5.3 shows the chip layout and the size of the test structure for each technology generation. Several ROs with identical circuit structures are included in all the test structures, which enables observation of variability trend over four technology generations.

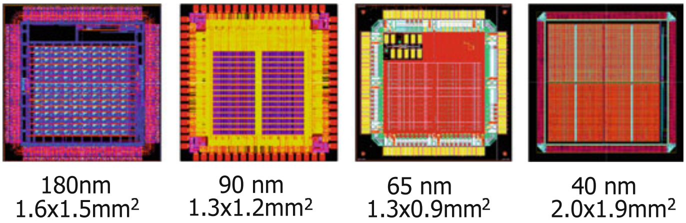


Fig. 5.3 RO (Ring Oscillator)-array test structure in four technology generations

5.1.4.2 Die-to-Die and Within-Die Variation

Table 5.1 lists the amount of D2D and WID components in oscillation frequency variations for 7-stage, 13-stage, 29-stage, and 59-stage inverter ROs. Standard deviations σ of oscillation frequencies normalized by their mean values μ are listed in percentile. The values of D2D components for each generation are almost the same regardless of the number of stages, although those values differ by technology generations. On the other hand, the value of WID components decreases as the number of stages increases. This happens due to the averaging effect of random variations as the number of stages increases.

Taking the 7-stage inverter RO as an example, we further decompose the WID variations into three components of Location-Specific, Across-Chip, and Random, where the location-specific component is a layout-dependent deterministic variation, the across-chip component is a gradually varying variation over the chip, and the random component is a random and uncorrelated variation over the chip [3]. Table 5.2 shows the amounts of standard deviations (σ/μ) for D2D and WID components and their breakdowns. It is clearly seen that the amount of WID random components rapidly increases as the technology scales. We estimate the amount of random variation for a single inverter from the stage-length dependency of random variations. The estimated value for a single inverter is also listed at the last row in Table 5.2. In the 40 nm process, a random variation with 7.5% standard deviation appears. It becomes clear that, in a scaled technology, we should take care of not only D2D variations but also WID variations in the design process.

Table 5.1 Comparison of WID (average) and D2D variations in σ/μ (%)

RO	180 (nm)		90 (nm)		65 (nm)		40 (nm)	
	D2D	WID	D2D	WID	D2D	WID	D2D	WID
INV7	4.6	1.5	3.2	1.5	0.9	1.7	2.0	2.4
INV13	4.3	1.2	3.2	1.2	0.9	1.4	2.2	1.8
INV19	4.1	1.1	3.2	1.1	1.0	1.3	2.2	1.5
INV29	4.2	1.0	3.2	1.0	1.0	1.1	2.2	1.3
INV59	–	–	–	–	0.9	1.1	2.0	1.0

Table 5.2 Variability breakdowns for 7-stage ROs

Variability component	Standard deviation σ/μ (%)			
	180 (nm)	90 (nm)	65 (nm)	40 (nm)
D2D	4.6	3.2	0.95	2.0
WID	1.5	1.5	1.7	2.4
Location specific	1.3	0.7	1.0	0.6
Across-chip	0.1	0.1	0.2	0.2
Random	0.6	1.4	1.3	2.3
A single gate	1.7	4.3	4.0	7.5

5.1.5 Variability Trend and Scaling Scenario

A major source of WID random variations is a discrete random dopant fluctuation (RDF). Due to the RDF, the threshold voltage V_{th} of a transistor fluctuates randomly. The amount of the fluctuation is proportional to the inverse of the square root of the channel area LW, which is expressed by the following equation.

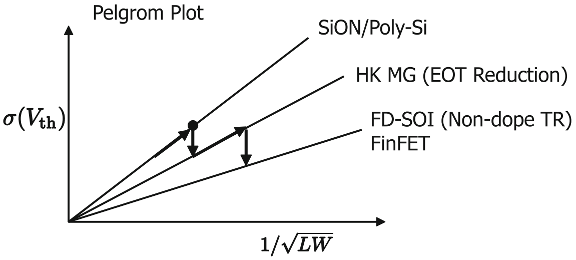
$$\sigma(V_{\mathrm{th}}) = \frac{A_{\mathrm{vt}}}{\sqrt{LW}}, \qquad A_{\mathrm{vt}} \propto t_{\mathrm{ox}} N_{\mathrm{A}}^{0.25}, \tag{5.1}$$

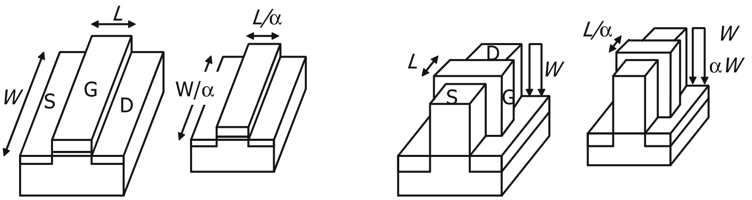
where L and W are the length and the width of the channel, respectively, and $t_{\mathrm{o}x}$ is the oxide thickness, and N_{A} is the dopant density of the channel. A graph that displays the amount of variation as a function of the -0.5-th power of the channel area is called a Pelgrom plot [4]. The gradient of the Pelgrom plot corresponds to A_{vt} which is proportional to the oxide thickness and the 0.25th power of the dopant density. It is, therefore, the shrink of the oxide thickness and the decrease in the dopant density lead to the reduction of variations.

In looking back the evolution of transistor structures in accordance with the progress in technology generations, up to around 45 nm technology nodes, polysilicon is commonly used for the gate material with SiON for the gate oxide material. After around 32 nm technology nodes, metal gates and oxide materials with high dielectric constant (HK MG: High-K Metal Gate) are introduced, which leads to the decrease in the EOT (Effective gate Oxide Thickness). Further, after around 22 nm technology nodes, fully depleted SOI (Silicon On Insulator) transistors and FinFETs that have zero or lightly doped channels are introduced. Those structural changes both contribute to the suppression of variations. Figure 5.4 shows the variability trend in accordance with the evolution of transistor structures. The amount of variations increases with technology scaling while the progress of transistor structures contributes to the abrupt reduction of variations. However, we also should be aware that the evolution of transistor structures introduces new sources of variation. For example in the case of FinFETs, performance variations due to the nonuniformity of metal-gate granularity become a big concern.

As shown in Sect. 5.1, the amount of random variations is related to device dimensions such as the area of the channel and the oxide thickness. The lower limit

Fig. 5.4 Variability trend and the evolution of transistor structures





	LP	HP
L	$1/\alpha$	$\leftarrow$
W	$1/\alpha$	$\leftarrow$
W/L	1	$\leftarrow$
LW	$1/\alpha^2$	$\leftarrow$
t_{ox}	$1/\sqrt{\alpha}$	$1/\alpha$
$\sigma(V_{\mathrm{th}})$	$\sqrt{\alpha}$	1

Increase

	LP	HP	LP	HP
L	$1/\alpha$	$\leftarrow$	$1/\alpha$	$\leftarrow$
W	α	$\leftarrow$	1	$\leftarrow$
W/L	α^2	$\leftarrow$	α	$\leftarrow$
LW	1	$\leftarrow$	$1/\alpha$	$\leftarrow$
t_{ox}	$1/\sqrt{\alpha}$	$1/\alpha$	$1/\sqrt{\alpha}$	$1/\alpha$
$\sigma(V_{\mathrm{th}})$	$1/\sqrt{\alpha}$	$1/\alpha$	1	$1/\sqrt{\alpha}$

Decrease

Fig. 5.5 Scaling scenarios for planar transistors and FinFETs

of the minimum supply voltage V_{ddmin}, which is defined as the minimum voltage that ensures a correct operation, is limited by the amount of variations in transistor characteristics [5]. Lowering the supply voltage for the reduction of power dissipation is essential for enabling a higher level of integration with technology scaling. It is therefore important to establish technology scaling that is compatible with variability reduction.

Figure 5.5 explains the scaling scenarios for planar transistors and FinFETs proposed by Itoh [5]. The left side of Fig. 5.5 shows the scaling scenarios of planar transistors for low-power (LP) applications and high-performance (HP) applications. It is shown that planar transistors cannot be scaled with reduced variations. On the other hand, scaling scenarios of FinFETs, in which the channel width (Fin height) is inversely scaled, are indicated in the second column in the right table of Fig. 5.5. Due to the inverse scaling of the channel width (Fin height), both of low-power (LP) FinFETs and high-performance (HP) FinFETs enable scaling with reduced variations [5]. The third column of the right table in Fig. 5.5 shows other possible scaling scenarios with a constant channel width (Fin height). The suppression of variability increase is achieved by the non-scaling of the channel width.

5.1.6 Section Summary

In this section, performance variability of MOS transistors due to technology scaling is overviewed. After explaining the classification and the sources of variations, observations of measured variations from 0.35 μm to 40 nm technologies are presented. In particular, it is shown that the amount of WID random variations increases rapidly with technology scaling. WID variations cannot be handled by a conventional worst-case design (corner-based design). It is important to develop countermeasure techniques that can consider the specific nature of each variation component.

5.2 Monitoring and Compensation for Variations in Device Characteristics

Hidetoshi Onodera, Kyoto University

5.2.1 On-chip Variability Monitoring and Compensation

Increased variability is an inherent issue associated with device scaling. The variability in device characteristics leads to the variability in circuit performance (“faults”), which may cause “errors” eventually resulting in malfunctions (“failures”). For ensuring higher yields, it is common to adopt the worst-case design method that assumes the device performance being located at the worst corner so that the circuit performance always meets specifications in the whole performance spread. On the other hand, a circuit that is designed under the worst corner inherently has an overhead in all aspects of speed, power, and area, except for the case that all the device variations are really located at their worst corners. Performance spread has been expanding especially for lower voltage operation. Based on a simulation assuming a model circuit in a 65 nm process, under the nominal supply voltage of 1.2 V, the spread in operating speed between the fast corner and the slow corner is 67%. When the supply voltage is decreased to 0.6 V, the performance spread becomes 200%. This means that huge overheads in power dissipation and area have to be compromised in order to guarantee the speed performance at the slow corner. The expanded performance spread associated with lower supply voltage becomes prominent even in WID (Within-Die) variations, as well as D2D (Die-to-Die) variations. According to a performance measurement of a NoC (Network-on-Chip) with regularly tiled 80 cores, it is reported that the variation of the maximum operating frequency $F_{\max}$ of each core is 28% under the nominal supply voltage of 1.2 V. It, however, expands to 62% for 0.8 V operation [6]. Besides D2D variations, WID location-specific variations should be considered for the target of variability modeling and compensation.

For coping with the performance variations, promising approaches include circuit-level techniques that mitigate the variations at the device level. An example is a method that monitors the delay of a critical-path replica [7]. Variations of device performances are evaluated by the delay time of the monitored path and the operating speed is controlled to the target value by adjusting the supply voltage and/or body (substrate) bias voltage. This method evaluates the variation by mapping both of nMOSFET variations and pMOSFET variations into a delay variation. It is, therefore, difficult to obtain the variation of each transistor independently. Even in a case where an nMOS transistor and a pMOS transistor vary in opposite directions, the same compensation has to be applied to both transistors. It, therefore, may happen that a leak current unnecessarily increases after compensation.

In order to overcome the variability issue, we have developed a method that monitors performance variations of an nMOSFET and a pMOSFET independently and compensates each performance by adjusting each body bias voltage. Applying this technique to a small region-by-region on a chip, we can compensate not only D2D variations but also WID variations.

This section is organized as follows. In Sect. 5.2.2, we explain the variability monitoring and compensation technique by localized body biasing. The developed circuit consists of a body bias generator and digital monitors that evaluate performance variations of an nMOSFET and a pMOSFET independently. A noticeable feature of the circuit is that it can be implemented in a cell-base design. The effectiveness of the proposed technique has been verified by a test chip fabricated in a 65 nm process. Details will be given in Sect. 5.2.3.

5.2.2 Variability Monitoring and Compensation by Localized Body Biasing

We have developed a variability compensation scheme that divides a chip into small regions and the variability of each region is compensated region-by-region [8]. After explaining the compensation scheme, variability monitoring circuits and a body bias generator are presented [8, 9].

5.2.2.1 Localized Body Biasing

In order to compensate D2D variations and WID variations as well, we divide the whole chip into small regions called "substrate islands" and variability monitoring and compensation of each region are performed by a self-monitoring and self-compensation circuit called SAM (Self-Adjustment Module). Figure 5.6 shows

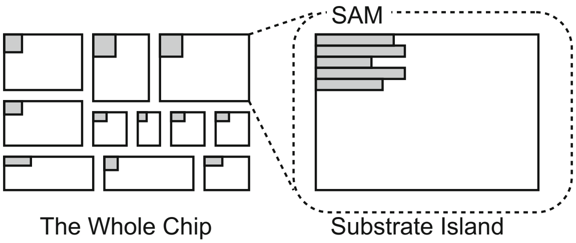


Fig. 5.6 Self-monitoring and self-compensation of transistor performance by localized body biasing

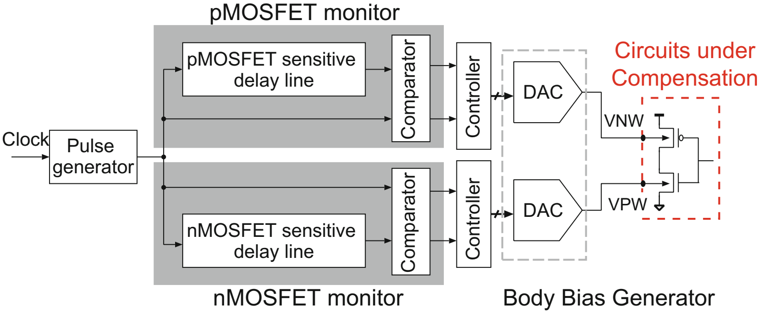


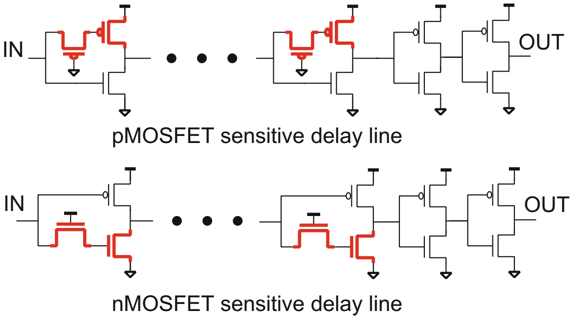
Fig. 5.7 Self-monitoring and self-compensation circuit

the compensation scheme by chip partitioning. Figure 5.7 illustrates the circuit configuration of the self-monitoring and self-compensation circuit SAM. It monitors performance variations of an nMOSFET and a pMOSFET independently by all-digital monitors. Based on the monitored results, a body bias generator supplies n-well and p-well bias voltages so that the performance of each type of transistors meets the target. The self-monitoring and self-compensation circuit (SAM) can be implemented in a cell-base design. Assuming the area of the substrate island is 0.1 mm^2, the area overhead of SAM is around 3% in our experiment. Due to its cell-base design, SAM can be integrated with a target circuit under compensation. By embedding SAM into white spaces of the target circuit, effective overhead can be further reduced.

5.2.2.2 Variability Monitoring

In order to measure performances of an nMOSFET and a pMOSFET independently, we have developed a monitor circuit sensitive only to an nMOSFET and a circuit sensitive only to a pMOSFET. Figure 5.8 shows the schematics of those circuits. Both circuits consist of path-transistor-inserted inverters followed by conventional

Fig. 5.8 Performance monitor circuits for pMOSFETs and nMOSFET



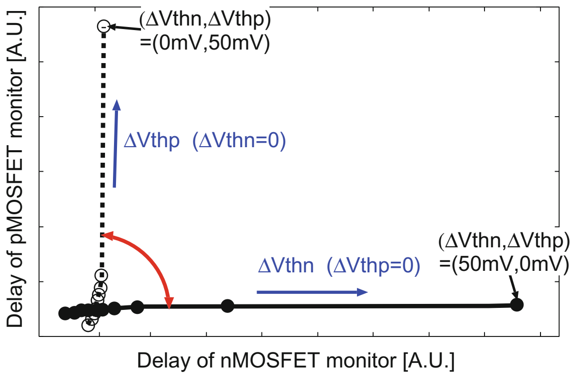


Fig. 5.9 Delay times of a pMOSFET monitor and a nMOSFET monitor when the threshold voltage of nMOSFET or pMOSFET is shifted

inverters. The path-transistor-inserted inverter in the upper circuit of Fig. 5.8 has a pMOS path-transistor inserted between the input of the inverter and the gate of the pMOS pull-up transistor, and it is sensitive only to the performance of pMOSFETs. The path-transistor-inserted inverter in the lower circuit of Fig. 5.8 has an nMOS path-transistor inserted between the input of the inverter and the gate of the nMOS pull-down transistor, and it is sensitive only to the performance of nMOSFETs.

The delay times of each monitor circuit are simulated and plotted in Fig. 5.9 by changing the threshold voltage of each type of transistor. The delay time of the nMOSFET monitor is sensitive to the threshold voltage change of nMOSFETs while it is not sensitive to that of pMOSFETs. The delay time of the pMOSFET monitor is sensitive only to the threshold voltage change of pMOSFETs. It is, therefore, possible to evaluate the performance variations of an nMOSFET and a pMOSFET independently from the measured delay time of each monitor circuit.

5.2.2.3 Variability Compensation by Adaptive Body Biasing

Based on the measured results of performance variations, the variations can be compensated by applying proper body (substrate) bias to each well. In this study, we have developed a body bias generator that supplies forward bias voltages so that performance compensation in the speeding-up direction can be possible. The circuit does not need an external voltage such that it generates body bias voltages only from a core voltage and a clock signal. Figure 5.10 shows the circuit topology. It consists of charge redistribution serial DACs (Digital-to-Analog Converters) of 6-bit accuracy and voltage followers by operational amplifiers. The circuit topology of the operational amplifier is shown in Fig. 5.11. For reducing power dissipation, a class-B output stage is applied. For enabling operation at the supply voltage of 0.6 V, the common mode level of input voltages is fixed to the half of the supply voltage.

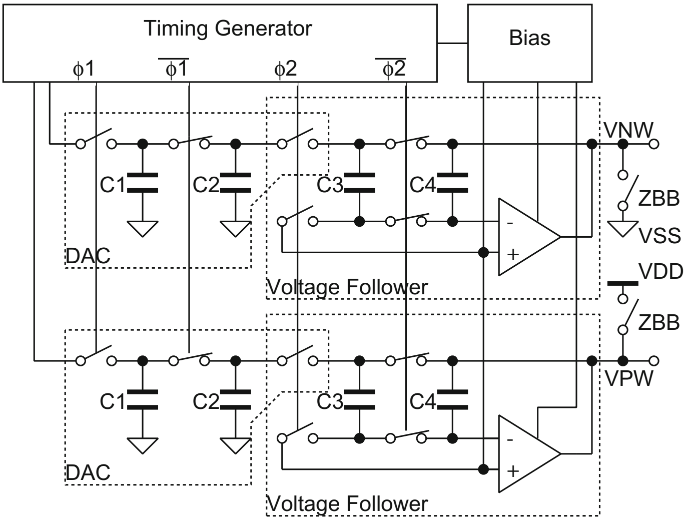


Fig. 5.10 Body bias generator circuit

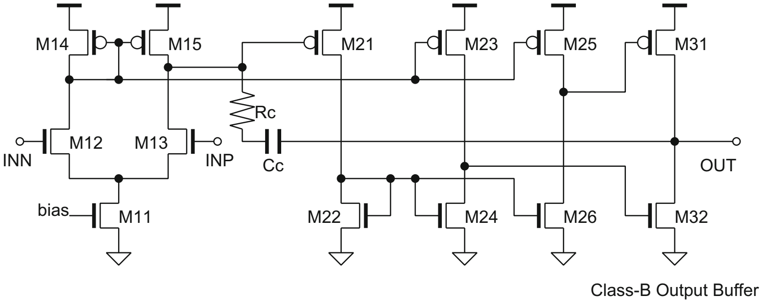


Fig. 5.11 Low-voltage and low-power operational amplifier with class-B output stage

When we apply a body bias voltage of up to 0.5 V, we can control the threshold voltage up to 100 mV assuming a 1/5 magnitude of drain current controllability compared to the gate. This amount is enough for compensating the device performance at the slow corner to the typical case, as shown in the next subsection.

5.2.3 Experimental Verification

The effectiveness of the proposed self-compensation scheme has been verified by a test circuit fabricated in a 65 nm process. Figure 5.12 shows a chip photograph of the test circuit together with a layout plot of the self-compensation circuit SAM.

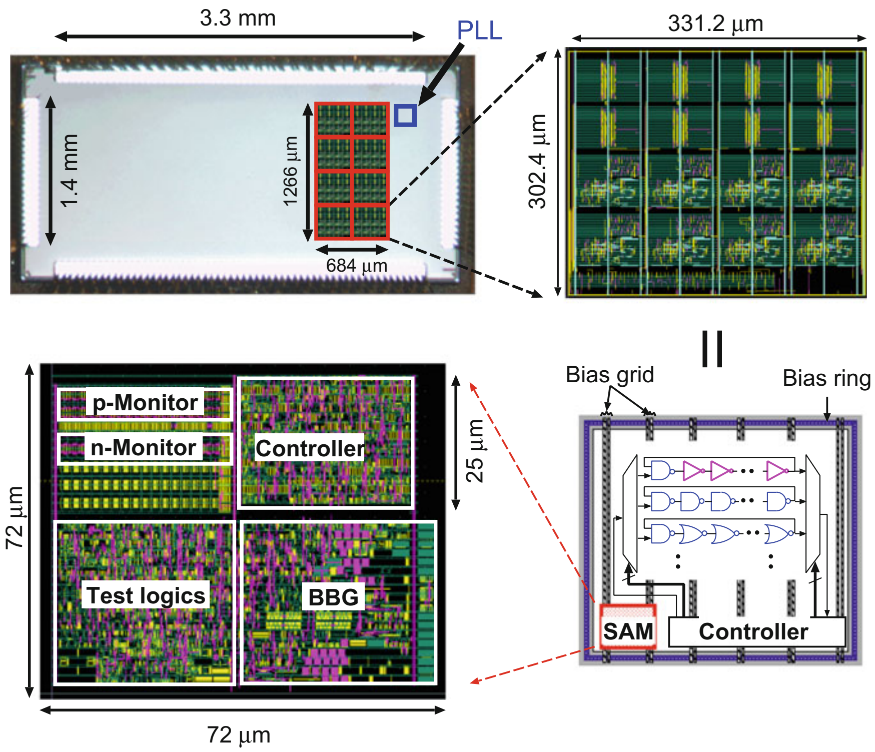


Fig. 5.12 Chip photograph and layout plot of a self-monitoring and self-compensation circuit

A substrate island of 0.1 mm^2 area (around 330 µm by 300 µm) is assumed. Eight substrate islands are integrated on the test chip. The self-monitoring and self-compensation circuit (SAM) is assembled in a 72 µm-by-72 µm space together with other logic gates. The total cell area of the self-monitoring and self-compensation circuit (SAM) is 2628 µm^2 which corresponds to the area overhead of 2.6%. Figure 5.13 shows the layout plot of the body bias generator. Colored cells are those that compose the body bias generator. They are integrated with other logic gates.

The test circuit has been fabricated under the typical condition (“TT”), and also under the four different process corners. Those corners correspond to four combinations of the fast corner (F) and the slow corner (S) for each type of transistor, and called “SS”, “FF”, “FS”, “SF,” respectively.

Operating speed of the test chip has been measured at the supply voltage of 0.7 V. It has been found that, except for the “FF” case, circuits in four other conditions do not meet the target speed without self-compensation. After enabling the self-measuring and self-compensation circuit (SAM), operating speed has been recovered in all the four conditions. Figure 5.14 shows the operating speeds of the nMOSFET monitor and the pMOSFET monitor before and after the self-compensation for “SS”, “TT”, “SF”, and “FS” chips. Generated body bias voltages are also indicated in the figure. In

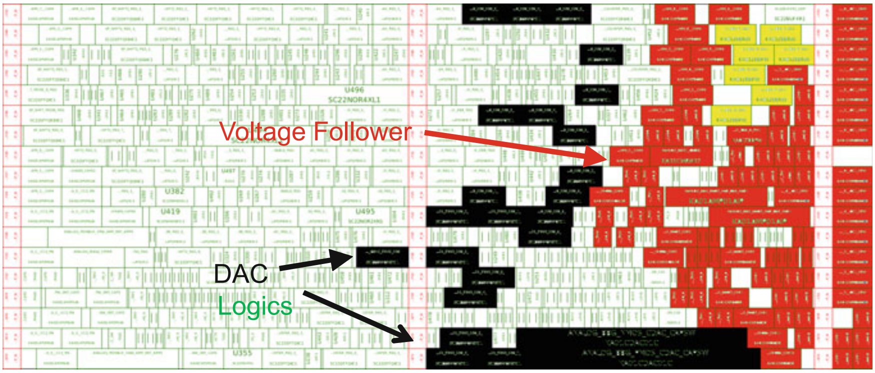
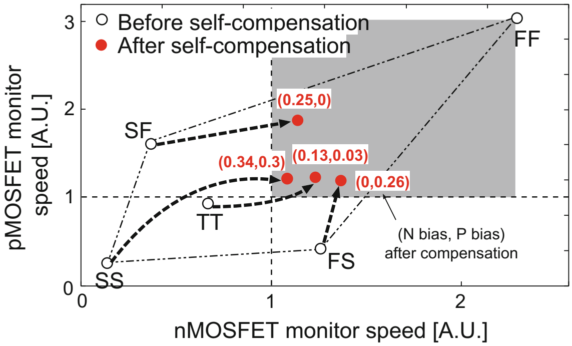


Fig. 5.13 Layout plot of a body bias generator (colored cells) integrated with other logic gates

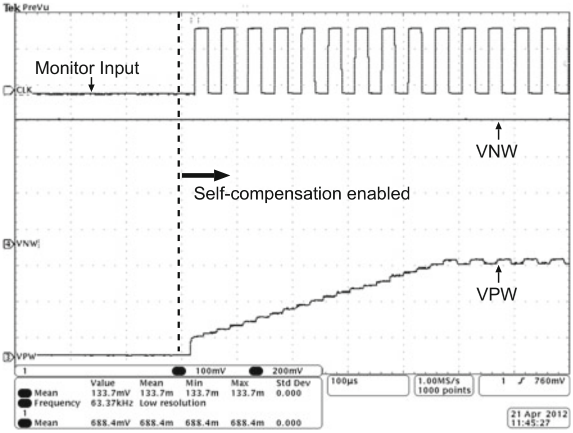
Fig. 5.14 Operating speeds of nMOSFET monitors and pMOSFET monitors before and after self-compensation for "SS", "TT", "SF", and "FS" chips



the cases of "FS" and "SF" chips, forward body bias is applied only to the body of the slow transistor. Figure 5.15 shows a transient response of the self-measuring and self-compensation circuit of an "SF" chip. Only the p-well voltage (VPW) for the slow nMOSFET ramps up until the target speed is recovered.

5.2.4 Section Summary

We have developed a variability monitoring and compensation scheme in which performance variations are self-monitored and self-compensated by body bias control so that the target speed is achieved. The whole chip is divided into a collection of small regions called "substrate islands," and each substrate island accommodates a self-monitoring and self-compensation circuit, thereby WID variations as well as D2D variations can be compensated. For performance monitoring, all-digital monitor circuits have been developed that can detect performance shifts of an nMOSFET and a pMOSFET independently. A body-bias generator that



```
              Value      Mean            Min       Max       Std Dev
● Mean        133.7mV    133.7m          133.7m    133.7m    0.000
● Frequency   63.37kHz   Low resolution
1
● Mean        688.4mV    688.4m          688.4m    688.4m    0.000
```

Fig. 5.15 Transient response of a self-measuring and self-compensation circuit of an "SF" chip

is compatible with cell-base design has been proposed. The proposed scheme has been verified and demonstrated by test circuits fabricated in a 65 nm process under five process corners of "TT", "SS", "FF", "SF", and "FS", where all the corner chips that do not meet the speed target have been successfully compensated to meet the speed goal.

5.3 Highly Accurate On-chip Measurement of Circuit Delay Time for Dependable VLSI Systems

Yukiya Miura, Tokyo Metropolitan University
Yasuo Sato, Kyushu Institute of Technology
Seiji Kajihara, Kyushu Institute of Technology

5.3.1 Purpose of Delay-Time Measurement

As semiconductors continue to be scaled down, process variation and circuit aging (degradation) affect the operation speed of the LSI [10, 11]. Variation and aging cause the change in the circuit delay time and result in a serious threat to LSI dependability. To enable a dependable design and preventive maintenance of a system, this section focuses on a method for measuring the delay time of the LSI accurately using an on-chip measurement circuit. To evaluate effects of variation and aging and to ensure correct operation of the system, the delay time (operation speed) of the LSI must be measured when the system is running. For the purpose, this section describes on-chip delay-time measurement that can measure the delay time of the LSI in the field where the LSI is used. In the on-chip measurement, an easy implementation and a small area overhead for the measurement circuit and a flexible path selection method for the delay-time measurement in the field as well as time resolution of the measurement circuit are needed. As the reasonable solution for satisfying those, a delay-time measurement method utilizing scan design is applied. In addition to variation and aging, the environmental parameters in the field such as temperature or voltage significantly affect the measured delay time. This section also describes a method to compensate for the effect of the environment on the measured delay time. This method can accurately evaluate the delay time due to variation and aging.

5.3.2 Overview of On-chip Delay-Time Measurement

Variation and aging affect the operation speed of the LSI. The LSI generally has enough margins for required specifications when it is designed, and it is shipped only if it has passed the production tests, including the at-speed test. In addition to the production tests, the burn-in test is applied to highly reliable products to reduce the occurrence of early failure. However, the actual delay time of each product is unknown even if it has passed the tests. Moreover, the effect of circuit aging becomes noticeable as the LSI continues to be used in the field, and the LSI will eventually become faulty (Fig. 5.16). Therefore, the delay time of the LSI in the field needs to be measured continuously to ensure a normal LSI performance and a long-term stable operation of the system.

To measure the delay time of the LSI with high time resolution, an on-chip circuit that measures the delay time of the LSI (e.g., the path delay time) is needed. Table 5.3 compares the main measurement methods using on-chip circuits.

A Vernier delay line (VDL) method measures the delay time using the difference between two buffers with different propagation delay times [12, 13]. A measurement circuit consists of two kinds of buffer chains and flip-flop chains. The time resolution of delay-time measurement by the VDL method is high because it uses the difference in propagation delay times between two buffer chains. However,

Fig. 5.16 Failure rate (bathtub curve)

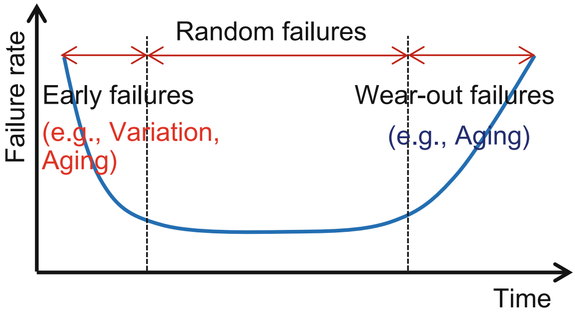


Table 5.3 Comparison of on-chip delay-time measurement methods

	VDL	OSC test	Razor/canary	Scan-based method
Circuit structure	Buffer and FF chains	Ring oscillator	Duplicated FF	Scan circuit
Measurement principal	Difference of delay time between two buffers	Self-oscillation period	Difference of sampling time between two FFs	Variable clock timing
Measurement path	Fix (depending on HW)	Fix (depending on HW)	Fix (depending on HW)	Flexibility (depending on TP)
Circuit overhead	Large (depending on # measuring paths)	Medium (depending on # measuring paths)	Large (depending on # measuring paths)	Small
Time resolution	High	High	Low	Medium

paths for delay-time measurement are fixed to the ones that have been assigned at the LSI design stage, and the amount of area overhead becomes larger as the range for measuring delay time becomes wider.

An oscillation (OSC) test method measures the delay time by using self-oscillation period of a path for delay-time measurement, where a ring oscillator is configured by connecting the input and output lines of the measuring path [14, 15]. From the configuration, time resolution of delay-time measurement is high and the measurement circuit is simple. However, paths for delay-time measurement must be fixed at the LSI design stage.

The Razor FF or the Canary FF uses a duplicated FF whose data sampling timings are slightly different [16, 17]. When values of two FFs are different, their methods result in detection of timing error. The delay time measured by those methods depends on the difference in sampling timings between two FFs. Moreover, the FF at the output of the path for delay-time measurement needs to be duplicated.

The above delay-time measurement methods are ad hoc techniques. On the other hand, for a structured technique, a scan-based delay-time measurement method has been proposed [18, 19]. This method measures the path delay-time of the circuit by scan design. In this method, by using variable clock timing, delay fault testing is applied to a path for delay-time measurement shortening a test clock period step by step. The method can prevent the increment of area overhead because it utilizes the scan architecture built in the chip. Moreover, it can select paths for delay-time measurement flexibly after the LSI is manufactured because the paths to be measured are determined by test patterns (TPs).

5.3.3 Delay-Time Measurement Using Scan Design

In a scan designed circuit, delay testing is applicable for a path (i.e., path under test (PUT)) between a scan-in FF and a scan-out FF (Fig. 5.17, upper). The PUTs are the group of paths that are sensitized by the applied test patterns. If the delay time of the path within the group exceeds a system clock period (i.e., at-speed), the delay fault of the path can be detected.

In a delay-time measurement method using scan design, variable launch clock timing or variable capture clock timing is used. Delay testing is applied to the PUT shortening the test clock period step by step. The clock period before the test is first failed is the actual delay time of the PUT (Fig. 5.17). In this method, since the delay test changing the test timing gradually is carried out repeatedly, the delay-time measurement itself takes time. Moreover, the accuracy of delay-time measurement depends on the resolution of the variable timing of the clock generation.

As a method for changing the period of the clock on the LSI chip, a variable clock generation method is usually used, such as the On-Die Clock Shrink (ODCS) technique [20, 21]. In this method, by inserting a buffer chain that can adjust the number of stages into a test clock line supplied to an FF, the generation timing of the test clock is made variable (Fig. 5.18). If the generation timing of the launch clock for scan-in FFs is delayed when scan testing is applied, the test clock period between the launch and capture clocks can be shortened. Therefore, the PUT can be tested by a clock period shorter than the system clock period (Fig. 5.18, right side). Note that, in this example, generation timing of the capture clock for scan-out FFs is fixed. The generation timing of the launch clock for the PUT (i.e., the path for delay-time measurement) is delayed gradually, and this test is carried out

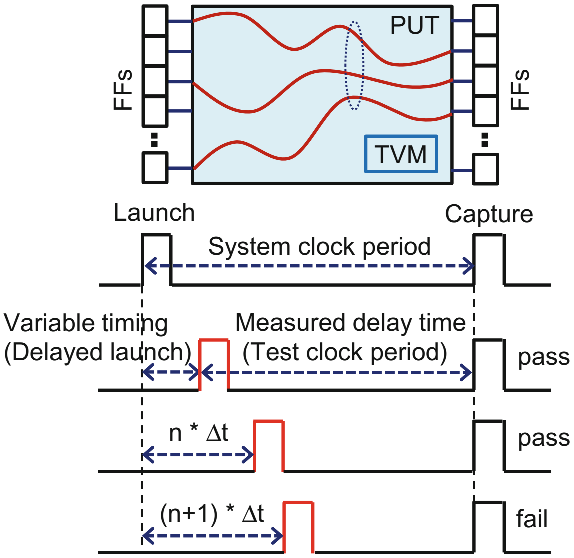


Fig. 5.17 Delay-time measurement by scan test

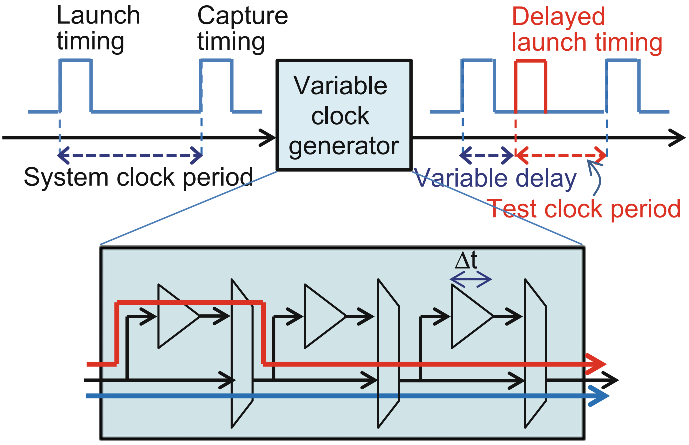


Fig. 5.18 Variable clock timing

repeatedly. When the PUT is identified as faulty (i.e., the test is failed), the timing margin of the path can be calculated as the value of (the system clock period)-(test clock period before first test failure). If the value of the minimum variable delay-time of the launch clock is Δt, as in the example in Fig. 5.17, the timing margin is at least $n * \Delta t$. This method can quantitatively measure the amount of the circuit delay caused by transistor variation and aging as a delay-time margin of a path. Here, note that time resolution for measuring the delay time depends on the change interval of variable clock timing. In addition, since PUTs depend on test patterns, paths for delay-time measurement can be selected after the LSI chip is manufactured. Thus, path selection is more flexible in this method than in other similar methods.

5.3.4 Delay-Time Measurement Considering Measuring Environment

Since delay time of a circuit depends on its operating environment, the measurement environment must be considered when delay time is measured in the field. This section introduces a method for highly accurate delay-time measurement using an aging detection technique in the field called dependable architecture with reliability testing (DART) technology [22]. The DART technology utilizes the function of the scan circuit in the Circuit Under Test (CUT). Delay time of the CUT is measured by using variable test clock timing.

The main purpose of the DART technology is delay-time measurement in the field for preventing delay-related errors (e.g., excessive circuit aging). To realize the field test, the operation environment of the CUT when the delay time is measured, which is

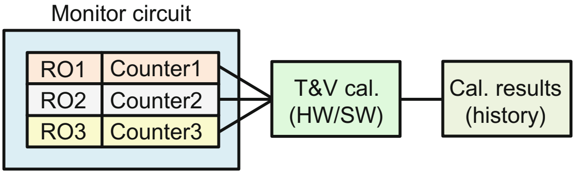


Fig. 5.19 Temperature and voltage monitor

the effect of a temperature and a voltage on the delay time, must be considered. For this purpose, a temperature and voltage monitor (TVM) is built in the CUT. The monitor circuit (sensor part) is a simple and small circuit consisting of three types of ring oscillators (ROs) with different frequency characteristics and counters [23] (Fig. 5.19). The monitor circuit is designed by a CMOS digital standard library. Temperature and voltage are concurrently estimated from RO frequencies (counter values) by fully digital processing, and measuring time is very short (several μs). The monitor circuit can be placed in plural locations and anywhere in a chip because of a small digital circuit. In an environment where temperature and voltage are not controlled, it is possible to compensate for the delay time while considering the measuring environment for the measured delay time by using the built-in TVM. From this technique, the correct timing margin of a circuit is measured. In addition, the technique can hold measured results of temperature and voltage as history data that can be utilized for estimating the busy condition of a chip in the field.

The DART technique has been applied to a circuit consisting of 7.2 M gates and 356 k FFs designed in 90-nm technology [22]. The DART circuit can be implemented with approximately 0.2% area overhead. Under the assumption of the error of the launch clock timing of 20 ps, accuracy of the whole delay-time measurement circuit is estimated as 27 ps by circuit simulation.

5.3.5 *Advantages of Delay-Time Measurement by the DART Technology*

Since the DART technology measures the delay margin of a circuit in a chip, it can quantitatively evaluate various factors during operation in the field after LSI shipment, such as variation and aging, which affect the circuit delay-time. This technology has the following advantages.

(1) Small area overhead utilizing existing scan circuits
(2) High accuracy of delay-time measurement depending on variable timing of the test clock
(3) Flexible path selection for delay-time measurement depending on test patterns
(4) Measurement of environment factors and compensation of measured delay time by using the temperature and voltage monitor.

The DART technology is considered suitable to be applied to the production test, the field test, the verification tool after manufacture (e.g., the tool for delay margin verification), the failure analysis tool, and analysis of chip use history (temperature and voltage log).

5.4 Timing-Error-Sensitive Flip-Flop for Error Prediction

Toshinori Sato, Fukuoka University
Ken Yano, Tokyo Institute of Technology[1]
Yuji Kunitake, Panasonic Corporation[2]

5.4.1 Timing-Error-Sensitive Flip-Flop

As semiconductor technologies are scaled, a new challenge of parameter variations has emerged. Process variation (P), supply voltage change (V), and temperature fluctuations (T) cause parameter variations (PVT variations) [24, 25]. PVT variations affect each transistor's threshold voltage, resulting in performance variations. Because each of these variations demands its own margin, the total design margins are always overestimated, resulting in wasteful increases in power consumption. In order to eliminate the wasteful power consumption, timing-error-sensitive FFs have been widely studied [16, 17, 26–31]. Because these FFs commonly require additional circuits to detect or to predict timing errors, however, the increase in the chip area, in turn, becomes a big concern. It might significantly enlarge area and power at the chip level. This section shows one possible solution.

First, in this subsection, a novel design philosophy, which is named typical-case design methodology, is introduced. Second, the concept of the timing-error-sensitive FFs, named Canary FF [17, 30], is explained. It is an essential component to make typical-case design practical. After that, an inevitable problem due to Canary FF is considered. Because a Canary FF cell is larger than the conventional D FF cell, it might have a serious impact on area and power in the chip level. And last, the related works are summarized.

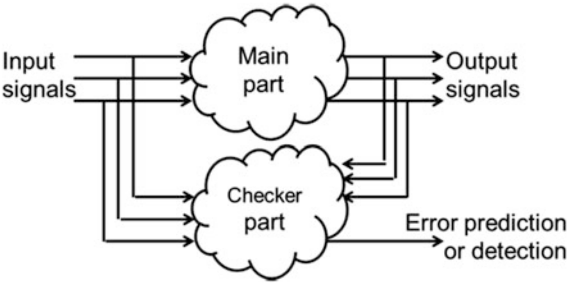
Typical-Case Design Methodology

The typical-case design methodology [32] addresses the overestimated design margin due to PVT variations by exploiting the observation that worst cases are rare. LSI designers can focus on typical cases, if there is an insurance mechanism for the worst cases. A single functionality is designed as two circuits. One is a performance-oriented circuit, where the correct functionality in the worst cases is

[1] Part of this work was done while the author was with Fukuoka University, Japan.

[2] Part of this work was done while the author was with Kyushu University, Japan.

Fig. 5.20 Typical case design



ignored. The other is a function-guaranteed circuit, where optimizing performance in the worst cases is not considered but the logical functions are guaranteed. Since designers should consider only one of the two severe constraints of performance and functionality, their design task becomes simple and easy.

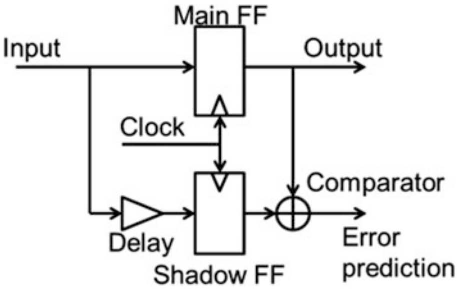
Figure 5.20 explains the concept of the typical-case design. Every critical function (for example, in performance or in power consumption) is designed as two circuits. They are named Main and Checker parts, respectively. Their functionality is equivalent but their roles and implementations are different with each other. The main part realizes the performance-oriented circuit, while the checker part realizes the function-guaranteed one. Hence, some errors might occur in the main part, and in such cases, the checker part supports the main part to guarantee their correct functionality. When an error is detected in the main part, the output of the main part is discarded and replaced with the one of the checker parts and then the state is recovered. In the other case where the error is predicted by the checker part, the possible error is avoided in the main part.

Canary Flip-Flop

In order to reduce the wasteful power, techniques combining Dynamic Voltage Scaling (DVS) with timing-error-sensitive FFs are studied [16, 17]. For example, because one technique decreases the supply voltage as low as possible without causing timing errors by predicting them on the fly, the useless power is eliminated. This section explains the Canary FF [17, 30] as a representative.

Figure 5.21 shows a Canary FF. It includes a redundant FF named shadow FF, a delay element, and a comparator. Due to the delay element, the shadow FF is more

Fig. 5.21 The Canary FF



vulnerable to timing errors than the main FF. If the values kept in the main and the shadow FFs do not match in the comparator, a timing error is predicted. In order to avoid the predicted error, DVS works to increase the supply voltage.

Area Overhead Problem

Because the Canary FF requires the additional circuit elements to detect errors, the area and power overheads in the LSI utilizing the Canary FF might be seriously large. From the preliminary study, it is found that a Canary FF cell is 2.5 times larger than the conventional D FF cell. In order to reduce the area overhead, a special type of scan FFs for production testing can be reused to realize a Canary FF [30]. Because they are already included in some LSIs, there is not any area overhead. However, unfortunately, all LSIs do not utilize special FFs due to cost consideration. In addition, the power overhead is not considered by this technique. Hence, the other solution required is a way to reduce the number of Canary FFs.

Related Works

A lot of timing-error-sensitive FFs are proposed [16, 17, 26–31]. The Razor FF [16] detects timing errors. It also has the shadow FF, where a delayed clock is delivered. In the case when the values kept in the main and the shadow FFs do not match, a timing error is detected. This technique is also applicable to detect soft errors [29]. NEC [28] utilizes the shadow FF to predict aging failures. Instead of delivering delayed clock to the shadow FF, a delay element is inserted between the previous logic stage and the shadow FF as a Canary FF. By detecting timing errors in the shadow FF, the main FF is protected. Agarwal et al. [26] propose a similar technique to predict defects. Intel extends their soft-error resilient FF [27] to support process variation diagnosis [31].

5.4.2 Selective Replacement Method

This subsection explains a technique to reduce the number of Canary FFs, presents its evaluation methodology, and shows evaluation results.

Replacement Strategy

In order to reduce the number of Canary FFs the distribution of path delay is investigated. Depending on the logical depth and wire length, each path has a different delay from the other ones. Timing errors will not occur on paths with small delays, even if PVT variations affect them. It is not necessary to replace any terminal D FFs, which are connected to the end of the short paths, with a Canary FF. Only timing-error-vulnerable FFs on the timing-critical paths should be replaced. This selective replacement method will reduce the chip-level overheads on area and power due to large Canary FF cells.

Figure 5.22 explains how the selective replacement method works [33]. There are three steps. The first step determines the target cycle time. A given RTL is

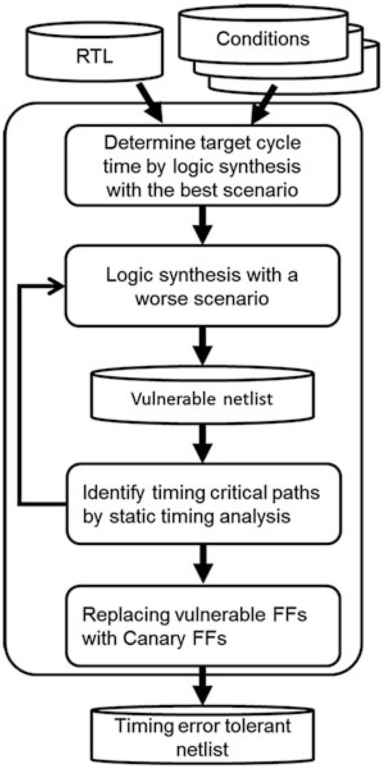

RTL
Conditions
Determine target cycle time by logic synthesis with the best scenario
Logic synthesis with a worse scenario
Vulnerable netlist
Identify timing critical paths by static timing analysis
Replacing vulnerable FFs with Canary FFs
Timing error tolerant netlist

Fig. 5.22 Selective replacement flow

logic-synthesized with the best-case scenario and the reported cycle time is used as the target cycle time. The second step identifies timing-critical paths, which are vulnerable to timing errors. For every worse condition, logic synthesis is performed and its netlist, which is vulnerable to timing errors at the target cycle time, is generated. Static timing analysis is performed on the netlist and the paths, which do not satisfy the target cycle time, are identified as timing-critical paths. This process is continued until all conditions are considered. The last step replaces terminal D FFs connected to the end of the timing-critical paths with Canary FF.

Evaluation Methodology

The selective replacement method is evaluated by two steps. First, a couple of available microprocessor cores are modified by adopting a few different designs. The purpose of the step is to evaluate the impact of Canary FFs on area and power at the chip level. Second, one of the processors is simulated in the instruction set level in order to evaluate how the wasteful power is reduced by DVS.

The first step goes as follows [34]. A tool for the design flow explained in the previous subsection is built [35]. It consists of SYNOPSYS's Design Compiler and an in-house Perl script. Using the tool, netlists of two processor cores, which are Toshiba's MeP [36] and an open-source miniMIPS [37], are generated. The

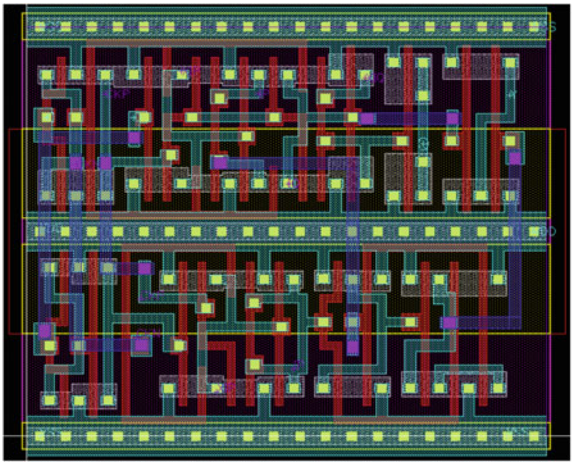


Fig. 5.23 Cell layout of Canary FF [34]

Table 5.4 Comparison between the Canary FF and D FF

	Canary FF	D FF
Area (um$^{2)}$	129.0	51.6
Power (uW)	63.0	25.0

standard cell library from Kyoto University [38] based on Rohm's 0.18 μm technology is used.

The following four netlists are designed.

1. Straightforward (S in Figs. 5.24 and 5.25): Logic synthesis (LS) and Place and Route (P&R) are performed with the typical-delay condition of the library. None of D FFs are replaced with Canary FFs. This design is vulnerable to timing errors and is impractical.
2. Worst-Case (W): LS and P&R are performed with the maximum-delay condition. None of D FFs are replaced with Canary FF. This is the traditional worst-case design result. It relies on a large timing margin so that it is protected from timing errors.
3. Canary (C): LS and P&R are performed with the typical-delay condition. The selective replacement method is used so that only timing-critical D FFs are replaced with Canary FF. This is the typical case design result. It is tolerable to timing errors with the help of the Canary FF.
4. All-Canary (A): LS and P&R are performed with the typical-delay condition and all D FFs are replaced with Canary FFs. While this is tolerable to timing errors, the increase in chip area and in power consumption will be serious.

The second step evaluates DVS based on instruction set simulations [39]. MeP simulator provided by Toshiba is used in order to generate execution traces. It is cycle accurate and models a quad-core processor. Benchmark programs are bubble, matmul, perm, qsort, queen, and sieve, which are selected from Stanford Integer

Table 5.5 Effect of selective replacement method

	Total # of FF (A)	# of Replaced FF (B)	Ratio (B/A) (%)
MeP	3,732	60	1.6
MiniMIPS	1,967	228	11.6

Benchmarks. A multiprogramming environment is assumed and thus the combinations of the programs running on the processor is ${}_6C_4 = 15$. Each trace is injected into an in-house simulator, which models DVS in detail.

Results

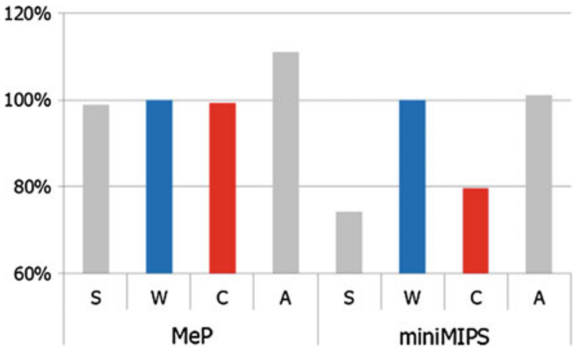
Figure 5.23 represents the hard macro cell of the Canary FF. Table 5.4 compares the cell with a D FF cell. It can be easily seen that both the cell area and the power consumption is 2.5 times larger in the Canary FF than in the D FF.

Table 5.5 presents the experimental results of the selective replacement method. The percentages of the Canary FFs over all FFs are 11.6% for miniMIPS and only 1.6% for MeP. The results confirm that the method works well to minimize the number of Canary FFs.

Figure 5.24 compares the chip areas of the four netlists after P&R. Each bar represents the relative chip area normalized by that of the Worst-Case case. In both processor cores, All-Canary is larger than Straightforward. This means that replacing all D FFs with Canary FF has the serious impact on the chip area. Even when All-Canary is compared with worst-case, both of which are timing-error-tolerant and practical, the difference between the two is not negligible. The selective replacement method successfully reduces the area overhead. The area is significantly reduced from All-Canary to Canary. When Canary is compared with Worst-Case, the chip area is reduced by 0.8% and 20.4% for MeP and miniMIPS, respectively. The reason why Canary is smaller than Worst-Case is that Canary FFs relax timing constraints and thus some of the powerful and large logical gates are not required.

Figure 5.25 summarizes the power consumption results. For each case, the bar is normalized by that of Worst-Case. It should be noted that the supply voltage is

Fig. 5.24 Chip area results



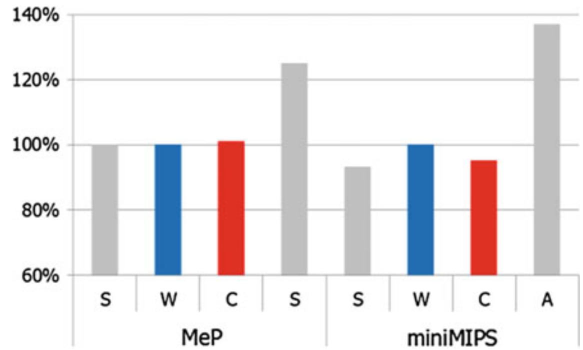


Fig. 5.25 Power consumption results

identical for every case. The selective replacement method works well also for limiting the increase in power consumption. It is considerably reduced from All-Canary to Canary. When Canary is compared with Worst-Case, the power consumption is slightly increased by 1% in the case of MeP and is rather decreased by 5% in the case of miniMIPS. From these results, Canary is superior to Worst-Case from the point of view of area and power.

The wasteful power is the difference between Worst-Case and Straightforward in Fig. 5.25. It is almost 0% for MeP and 7% for miniMIPS. This does not mean the typical-case design is useless. Please remember that the supply voltage is identical for both cases. If DVS was applied to Canary, power consumption would be further decreased. This is possible because Canary FF predicts timing errors and thus the supply voltage can be decreased without causing timing errors. This is evaluated in the second step. The instruction set simulation results of a quad-core MeP show that applying DVS reduces energy consumption by 21.2% and 18.6% on average without and with considering process variation, respectively [39]. The impact on performance is very small and the degradation is less than 2%.

In summary, the typical-case design successfully reduces power consumption without serious impact on chip area or on performance.

5.4.3 Conclusions

This chapter introduces the typical case design with the help of the Canary FF.

Although the Canary FF cell is 2.5 times larger than the conventional D FF, the selective replacement method minimizes the increase in area and power at the chip level. In the case of the single-core MeP, only 2% of D FFs are replaced. The chip-level P&R results show that there are no significant differences in design quality between the traditional worst case design and the typical case design. If DVS is applied to the quad-core MeP, energy consumption is reduced by 18.6% without serious impact on performance. The results show that the timing-error-tolerant design utilizing the Canary FF is practical and is one of the promising design methods for future process technologies vulnerable to PVT variations.

5.5 Fine-Grain Assist Bias Control for Dependable SRAM

Koji Nii, Renesas Electronics Corporation

5.5.1 Introduction

For low-power and low-voltage operation below 1.0 V, robust design under process variations is necessary to produce a deep-submicron dependable system-on-chip (SoC). Especially, embedded SRAMs are facing scaling limitations because of increasing V_{th} variation of transistors. To date, many design techniques that introduce SRAM assist circuits to enhance the read/write-margin have been reported [40–46]. These techniques are useful for maximizing the operating margins by controlling the bias of wordlines (WLs), bitlines (BLs), and power supply source lines of bitcells (VDM). These reports are discussed using DC characteristics [47, 48] later in the next subsection. To further improve read/write-margin, some recent assist circuits use the benefits of dynamic stability [49–53]. By using self-adjustable circuits or trimming variable elements with a fuse by memory BIST, each bias is adjusted automatically to optimum bias depending on the PVT (process-originated V_{th}) variations. However, since these approaches are introduced in a unit of large memory macro, the improvement of the minimum operating voltage (V_{min}) of SRAM is smaller or even worse than expected with the increasing number of SRAM capacity. Therefore, an individual bias control for each WL, BL, and VDM is necessary for additional improvement of SRAM V_{min}. As described herein, we propose a fine-grained assist bias control technique for enhancing the read/write-margin under low supply voltage operation: less than 1.0 V.

5.5.2 Conflicting Issues of Read-Assist

Figure 5.26 shows a typical schematic of the 6T SRAM bitcell with read- and write-assist circuits. Here, the WL bias lowering technique as a read-assist circuit is introduced to enhance the read-margin. In a write-operation, the VDM bias of the selected column is lowered using a write-assist circuit to improve the write-margin. Figure 5.27a portrays how the butterfly curves are affected by local V_{th} variations when the WL is activated. The static noise margin (SNM) [47] is improved by lowering the WL voltage as a read-assist. Figure 5.27b also shows the

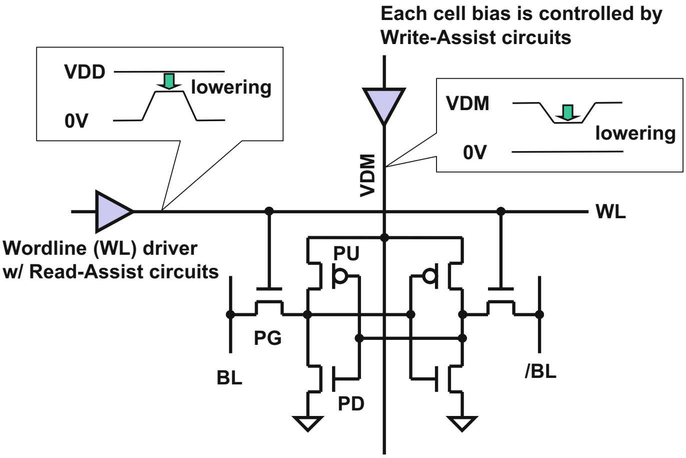


Fig. 5.26 Typical scheme of 6T-SRAM bitcell with read-/write-assist circuits

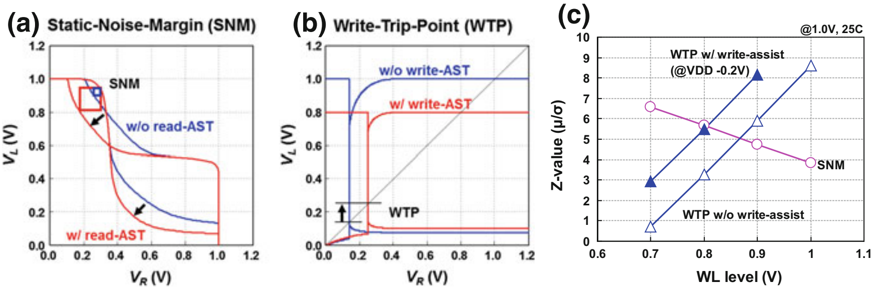


Fig. 5.27 **a** Static noise margin (SNM) [47] of the 6T SRAM w/ and w/o read-assist. **b** Write-trip-point (WTP) [48] of the 6T SRAM w/ and w/o write-assist. **c** SNM and write margin (WTP) of the 6T SRAM depending on WL bias lowering

write-margin, defined as the write-trip-point (WTP) [48] with consideration of local V_{th} variations. Because of the VDM bias lowering in a selected column as a write-assist, the WTP voltage becomes higher, thereby improving the write-margin. Figure 5.27c presents the dependences of SNM and WTP on WL bias. The typical supply voltage (VDD) is 1.0 V and the temperature is 25 °C. The Z-value of the vertical axis is defined by the mean value (μ) over the standard deviation (σ), indicating robustness against V_{th} local variations of the bitcell. Although the WL lowering improves the SNM, the write-margin degrades greatly below 0.8 V, even if the write-assist is introduced. For this reason, the bitcells with smaller write-margin are deteriorated by the WL lowering because of decreased overdrive voltage Vgs of the pass-gate (PG).

5.5.3 Concept of a Fine-Grained Assist Control

Monte Carlo simulation incorporating local V_{th} variations shows that the read-margin (denoted as SNM) and write-margin (denoted as WTP) are negatively correlated, as portrayed in Fig. 5.28. Note that there is no appearance probability of both read- and write-failure. In addition, read-margin-less bits have sufficient margin for write-operation, whereas write-margin-less bits have sufficient margin for a read operation. Ideally, the WL bias lowering is only introduced for the read-margin-less bitcells; however, that is not feasible because each WL is commonly connected to the bitcells arranged in the same row of the cell array. Thus, we propose the individual assist bias control for a fine-grained cell array region with feasibility. Figure 5.29 presents the concept of the proposed fine-grained assist bias control. A number of rows are grouped as an X-segment for read-assist to suppress the WL bias commonly. Otherwise, several columns are grouped as a Y-segment for write-assist to suppress the cell VDD bias commonly. Each read-assist bias corresponding to X-segment is controlled individually where the read-failure bits exist or not. Each write-assist bias corresponding to the Y-segment is also controlled individually where the write-failure bits exist or not.

5.5.4 Practical Dependable SRAM Macro with Fine-Grained Assist Control

Figure 5.30 shows a schematic diagram of the proposed 128 kb SRAM macro. For read-assist operation, we divided the 256 × 512 cell array into 16 X-segments and 8 Y-segments. Each X-segment has 256 columns by 32 rows; each Y-segment has 32 columns by 512 rows. We assign the two digits as bias conditions in each segment so that the assist bias can be selected with four levels. Consequently, additional 64 registers in all are necessary to set the bias conditions for read-assist and write-assist. A practical read-assist circuit and write-assist circuit are presented

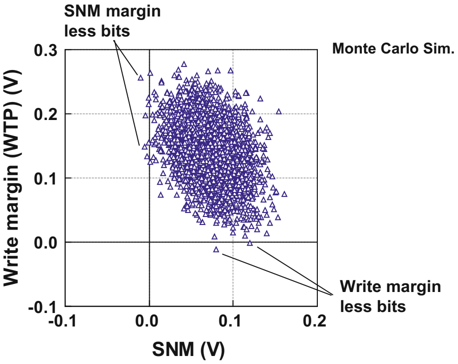


Fig. 5.28 Correlation between SNM and WTP using Monte Carlo simulation for a 6T SRAM bitcell

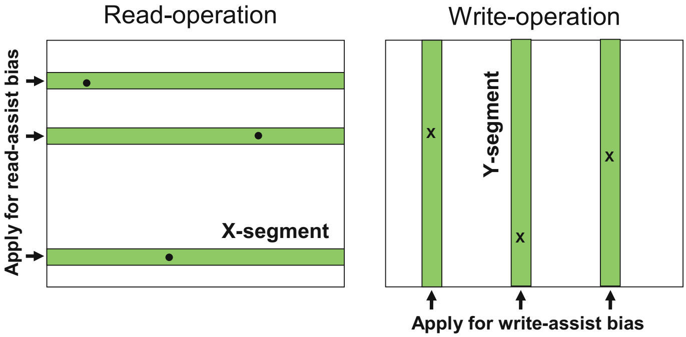


Fig. 5.29 Concept of a fine-grained assist bias control

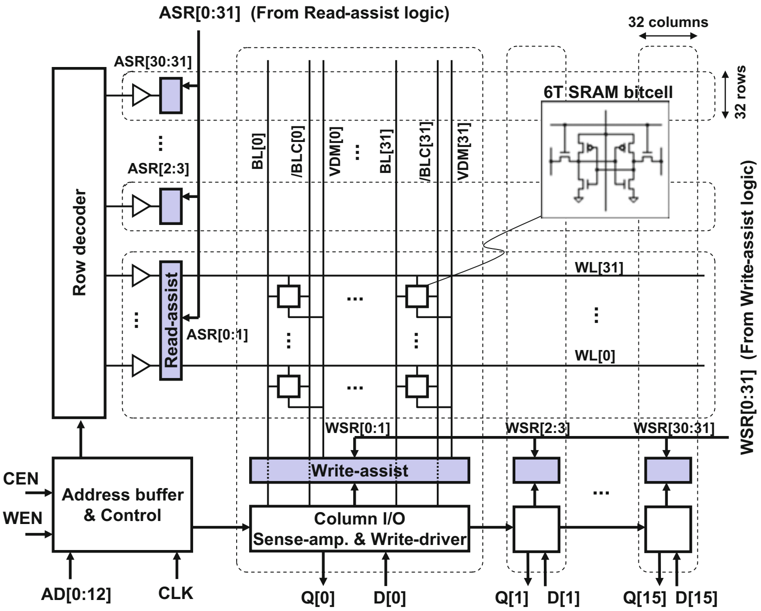
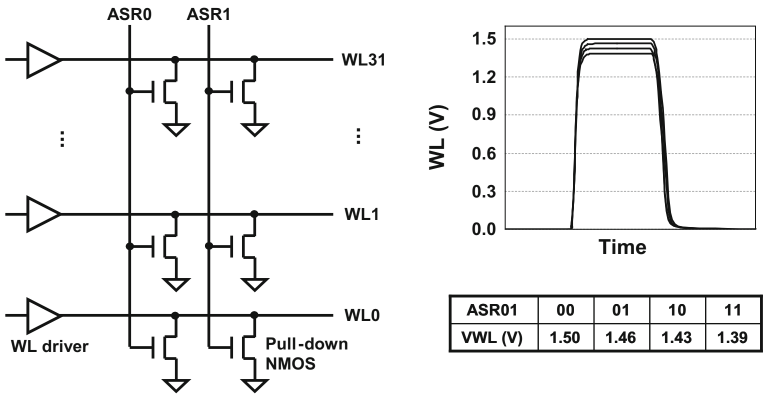


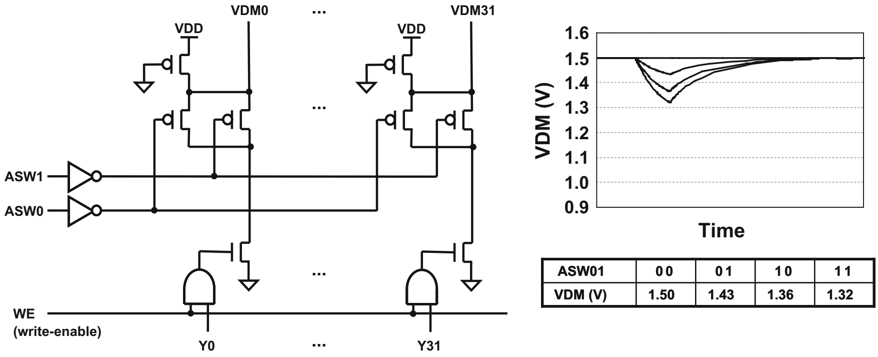
Fig. 5.30 Schematic diagram of proposed SRAM macro

in Fig. 5.31 and Fig. 5.32, respectively. The read-assist circuit has two pull-down NMOS in each row, whose drain and gate are connected to each WL and assist signal: ASR0, ASR1. The biases from WL0 to WL31 have much lower voltage to enhance the SNM if the ASR0 = ASR1 = “H”, reducing by 110 mV as shown in the simulated waveform. On the other hand, when the ASR0 = ASR1 = “L”, they are equal to VDD levels. In such cases, all bitcells within an X-segment have much margin for SNM and originally need not enhance the read-margin any further. Each column has two pull-down NMOS, whose gates are connected to the assist signals



ASR01	00	01	10	11
VWL (V)	1.50	1.46	1.43	1.39

Fig. 5.31 Proposed practical read-assist circuit and its simulated waveform



ASW01	00	01	10	11
VDM (V)	1.50	1.43	1.36	1.32

Fig. 5.32 Proposed practical write-assist circuit and its simulated waveform

ASW0 and ASW1 in Fig. 5.32, respectively. In the write-operation, the write-enable WE and one of the column decode signals from Y0 to Y31 are activated, turning on the footer NMOS corresponding to the selected column. If the assist signals ASW0 and ASW1 are equal to “H”, then both pull-down PMOS turn on. Then the much lower bias for VDM is driven to enhance the write-margin. Otherwise, if the ASW0 and ASW1 are equal to “L”, then VDM remains at a constant level as a supply voltage of VDD. In this case, all bitcells within the Y-segments have good write-margins. There is no need to apply the assist bias further. Each VDM level according to the ASW0 and ASW1 is also shown in Fig. 5.32.

Figure 5.33a presents the proposed dependable embedded SRAM with assist logics and a memory BIST. To control the individual bias, additional registers, which store the bias conditions, are added to the assist logics. Figure 5.33b shows that these registers are set by an assist controller according to the test flows. After

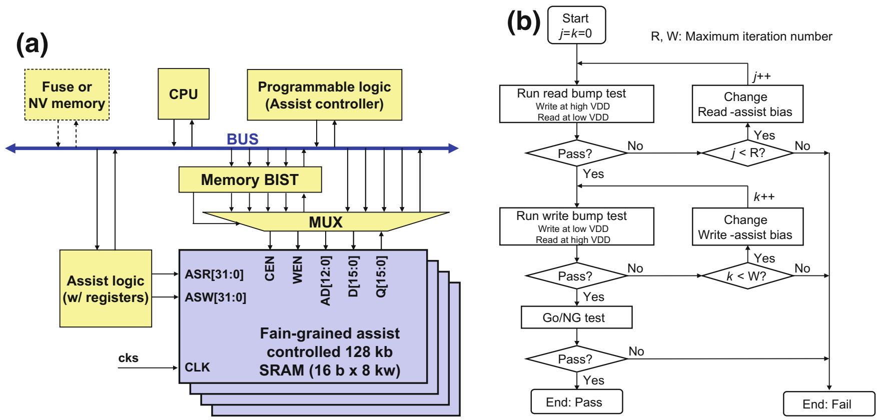


Fig. 5.33 **a** Proposed dependable SRAM with an assist logic and a memory BIST. **b** Test flows of the proposed fine-grained assist bias control with memory BIST

the screening test, the values of registers of the assist logics are stored to the fuse elements located in the same die or to the external nonvolatile memory. In the field, the registers of the assist logic are set merely by loading data from the fuse blocks or external nonvolatile memory at power-on. The assist bias can be turned in the field by running a diagonal memory BIST if a nonvolatile memory is used. This contributes to the improvement of reliability against aging degradation of the drain currents such as NBTI of pull-up PMOS in bitcells.

5.5.5 90 nm Test Chip Implementation and Measurement Results

To evaluate the effect of proposed assist circuits, we design and fabricate micro-controller test chips using 90-nm Low-Standby Power (LSTP) CMOS technology. Figure 5.34a shows a microphotograph and layout plot of the test chip. The die size is 59.3 mm^2. The test chip has a CPU with peripheral logics instruction memories, and an embedded Programmable Logic matrix (ePLX) [54]. It also includes 40 instances of 128 kb proposed SRAM macros, totally embedding 5-Mb storage capability. Memory BIST circuits are also implemented for screening the failure bits. During the setup period after power-on, the ePLX roles as an assist controller to set each assist bias condition to the registers of assist logics, temporarily. These assist logics with registers are placed around the SRAM macros. For the entire embedded 5 Mb SRAM, the logic gate counts including the memory BIST and assist logics are 244 k-gates. Although the overhead of the additional assist logic is 63% of the total logic for 5 Mb SRAM, the area penalty is less than 1% of the die area. Figure 5.34b portrays the layout plot of the proposed 128 kb

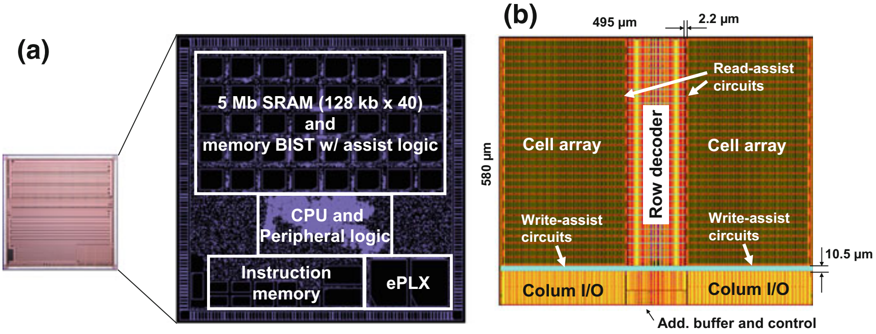


Fig. 5.34 **a** Microphotograph and layout plot of designed and fabricated test chips using 90-nm CMOS technology. **b** Layout plot of the 128 kb SRAM macro with fine-grained R/W assist bias control

Table 5.6 Features of the fabricated test chips

Technology	90-nm LSTP CMOS bulk process with 6 Cu-metals and AL-top-metal
Chip size	7.7 mm × 7.7 mm
Target speed	150 MHz @1.5 V ± 10%, −40 to 125 °C
Key IPs	
(1) Programmable logic	Programmable logic matrix (ePLX) [54]
(2) Embedded SRAM	5 Mb (128 kb × 40 instances)
	128 kb macro size: 580 µm × 495 µm
	6T bitcell size: 1.25 µm^2
(3) Memory BIST w/ Peri. and SRAM assist logic	244 k gates
	(BIST + Peri: 150 k gates, Assist: 94 k gates)

SRAM macro. The fine-grained read-assist circuits are placed between cell array and row decoder, whereas the fine-grained write-assist circuits are inserted between the cell array and column I/O peripheral blocks. The macro is 580 µm × 495 µm; the area overheads of the assist circuits are only 3%. The test chip features are presented in Table 5.6.

Figure 5.35a portrays the measured typical fail-bit-maps (FBM) of 25.5 kb (128 kb × 2) SRAM at 0.7 V and 25 °C. We observed each address of failure bits for read-bump tests; write-bump tests were different from each other. Applying an individual bias for each X-/Y-segment including failure bits enables successful 0.7 V operation. Figure 5.35b shows the measured V_{min} of 1 Mb SRAM for each case: original without assists, conventional assists, and proposed assists. It sometimes happens that the V_{min} of conventional assists becomes worse than the original without assists because all WLs are suppressed to improve the read-margin first

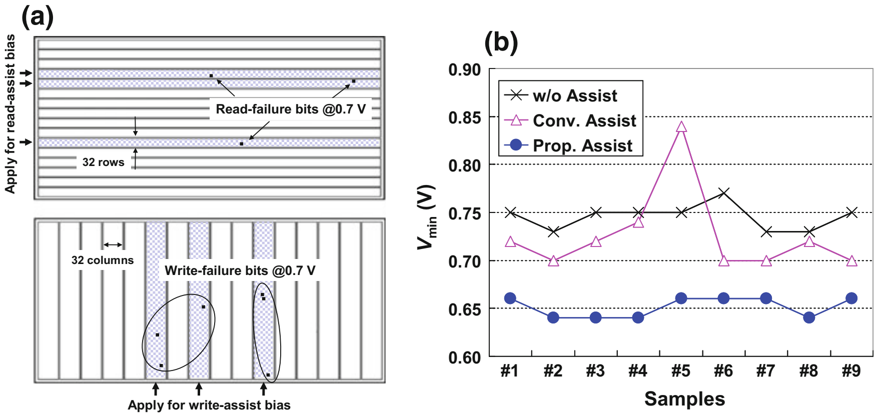


Fig. 5.35 **a** Measured fail-bit-maps (FBMs) of 256 kb (128 kb × 2) SRAM for read-/write-bump tests. **b** Measured $V_{\min}$ of 1 Mb SRAM

despite the existing write-margin less bits. In this conventional case, the write-margin gets worse adversely, resulting in the $V_{\min}$ degradation. Although the proposed fine-grained assist bias control technique suppresses the WL bias for some segments, resulting in further $V_{\min}$ improvement, which is from 70 to 110 mV compared to the original one. The results show that the minimum $V_{\min}$ of 1 Mb SRAM is achieved as 0.64 V, which is improved by 21% at most compared to the conventional assist circuits.

5.5.6 *Summary*

In this section, a fine-grained assist bias control technique for enhancing read-/ write-margins of an embedded SRAM [55] is introduced. Further improvement of $V_{\min}$ of SRAM macro was presented with a small area overhead. We designed and fabricated test chips with plural 128 kb SRAMs using 90 nm CMOS technology. The evaluation results demonstrated that $V_{\min}$ was 0.64 V, which is 21% better than that achieved using conventional techniques.

Acknowledgments This work was supported in part by JST CREST DVLSI project, by JSPS Grant-in-Aid for Exploratory Research, and by the fund from Central Research Institute of Fukuoka University. It was also supported by VDEC of the University of Tokyo in collaboration with Synopsys Inc., Cadence Design Systems Inc., and Rohm Co. Ltd. The cell library used in this research was developed by Tamaru and Onodera laboratory, Kyoto University and is released by Kazutoshi Kobayashi of Kyoto Institute of Technology. The authors sincerely appreciate Shunitsu Kohara of Toshiba Corporation for helping them use MeP simulator.

References

1. J.K. Kuhn et al., Process technology variation. IEEE Trans. Electron Devices **58**(8), 2197–2208 (2011)
2. H. Onodera, Variability modeling and impact on design, in *Proceedings of IEDM*, pp. 701–704, Dec 2008
3. H. Onodera, H. Terada, Characterization of WID delay variability using RO-array test structure, in *Proceedings of ASICON,* pp. 658–661, Oct 2009
4. M. Pelgrom et al., Matching properties of MOS transistors. IEEE J. Solid-State Circ. **24**(5), 1433–1440 (1989)
5. K. Itoh, Adaptive circuits for the 0.5-V nanoscale CMOS era, in *IEEE ISSCC Digest of Technical Papers,* pp. 14–20, Feb 2009
6. S. Dighe et al., Within-die variation-aware dynamic-voltage-frequency-scaling with optimal core allocation and thread hopping for the 80-core teraFLOPS processor. IEEE J. Solid-State Circ. **46**(1), 184–193 (2011)
7. M. Floyd et al., Introducing the adaptive energy management features of the power 7 chip. IEEE Micro **21**(2), 60–75 (2011)
8. N. Kamae et al., A body bias generator compatible with cell-based design flow for within-die variability compensation, in *Proceedings of ASSCC*, pp. 389–392, Nov 2012
9. A built-in self-adjustment scheme with adaptive body bias using P/N-sensitive digital monitor circuits. in *Proceedings of ASSCC*, pp. 101–104, Nov 2012
10. X. Lu, Z. Li, W. Qiu, D.M.H. Walker, W. Shi, PARADE: parametric delay evaluation under process variation, in *Proceedings of International Symposium on Quality Electronic Design*, pp. 276–280, Mar 2004
11. W. Wang, V. Reddy, A.T. Krishnan, R. Vattikonda, S. Krishnan, Y. Cao, Compact modeling and simulation of circuit reliability for 65-nm CMOS technology. IEEE Trans. Device Mater. Reliab. **7**(4), 509–517 (2007)
12. T.E. Rahkonen, J.T. Kostamovaar, The use of stabilized CMOS delay lines for the digitization of short time intervals. IEEE J. Solid-State Circ. **28**(8), 887–894 (1993)
13. R. Datta, A. Sebastine, A. Raghunathan, J.A. Abraham, On-chip delay measurement for silicon debug, in *Proceedings of Great Lakes Symposium of VLSI*, pp. 145–148, Apr 2004
14. K. Arabi, H. Ihs, C. Dufaza, B. Kaminska, Dynamic digital integrated circuit testing using oscillation-test method. Electron. Lett. **34**(4), 762–764 (1998)
15. X. Wang, M. Tehranipoor, R. Datta, Path-RO: a novel on-chip critical path delay measurement under process variations, in *Proceedings of International Conference on Computer-Aided Design*, pp. 640–646, Nov 2008
16. D. Ernst, N.S. Kim, S. Das, S. Pant, T. Pham, R. Rao, C. Ziesler, D. Blaauw, T. Austin, T. Mudge, K. Flautner, Razor: a low-power pipeline based on circuit-level timing speculation, in *Proceedings International Symposium on Microarchitecture*, pp. 7–18, Dec 2003
17. T. Sato, Y. Kunitake, A simple flip-flop circuit for typical-case designs for DFM, in *Proceedings of International Symposium on Quality Electronic Design*, pp. 539–544, Mar 2007
18. B.I. Dervisoglu, G.E. Stong, Design for testability: using scan path techniques for path-delay test and measurement, in *Proceedings of International Test Conference*, pp. 365–374, Oct 1991
19. S. Jin, Y. Han, H. Li, X. Li, Unified capture scheme for small delay defect detection and aging prediction. IEEE Trans. Very Large Scale Integr. Syst. **21**(5), 821–833 (2013)
20. S. Tam, S. Rusu, U.N. Desai, R. Kim, J. Zhang, I. Young, Clock generation and distribution for the first IA-64 microprocessor. IEEE J. Solid-State Circ. **35**(11), 1545–1552 (2000)
21. T. Xanthopoulos (ed.), *Clocking in Modern VLSI Systems* (Springer, New York, 2009)
22. Y. Sato, S. Kajihara, T. Yoneda, K. Hatayama, M. Inoue, Y. Miura, S. Ohtake, T. Hasegawa, M. Sato, K. Shimamura, DART: dependable VLSI test architecture and its implementation, in *Proceedings of International Test Conference*, 15.2, Nov 2012

23. Y. Miura, Y. Sato, Y. Miyake, S. Kajihara, On-chip temperature and voltage measurement for field testing, in *Proceedings of European Test Symposium*, p. 181, May 2012
24. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi V. De, Parameter variations and impact on circuits and microarchitecture, in *40th Design Automation Conference*, pp. 338–342, June 2003
25. O. Unsal, J. Tschanz, K. Bowman, V. De, X. Vera, A. Gonzales, O. Ergin, Impact of parameter variations on circuits and microarchitecture. IEEE Micro **26**(6), 30–39 (2006)
26. M. Agarwal, B.C. Paul, M. Zhang, S. Mitra, Circuit failure prediction and its application to transistor aging, in *25th VLSI Test Symposium*, pp. 227–286, May 2007
27. S. Mitra, N. Seifert, M. Zhang, Q. Shi, K.S. Kim, Robust system design with built-in soft-error resilience. IEEE Comput. **38**(2), 43–52 (2005)
28. T. Nakura, K. Nose, M. Mizuno: Fine-grain redundant logic using defect-prediction flip-flops, in *IEEE ISSCC Digest of Technical Papers*, pp. 402–402, Feb 2007
29. M. Nicolaidis, Time redundancy based soft-error tolerance to rescue nanometer technologies, in *17th VLSI Test Symposium*, pp. 86–94, Apr 1999
30. T. Sato, Y. Kunitake, Canary: a variation resilient ff to eliminate design margin for energy reduction. IPSJ J. **49**(6), 2029–2042 (2008) (in Japanese)
31. M. Zhang, T.M. Mak, J. Tschanz, K.S. Kim, N. Seifert, D. Lu, Design for resilience to soft errors and variations, in *13th International On-Line Testing Symposium*, pp. 23–28, July 2007
32. T. Sato, I. Arita, in *Constructive Timing Violation for Improving Energy Efficiency*, ed. by L. Benini, M. Kandemir, J. Ramanujam. Compilers and Operating Systems for Low Power (Springer, 2003), pp. 137–153
33. Y. Kunitake, T. Sato, H. Yasuura, T. Hayashida, A selective replacement method for timing-error-predicting flip-flops. J. Circ. Syst. Comput. **21**(6), 14 (2012)
34. K. Yano, T. Hayashida, T. Sato, Improving timing error tolerance without impact on chip area and power consumptions, in *15th International Symposium on Quality Electronic Design*, pp. 389–394, Mar 2013
35. K. Yano, T. Yoshiki, T. Hayashida, T. Sato, An automated design approach of dependable VLSI using improved Canary FF, in *7th International Workshop on Unique Chips and Systems*, pp. 34–39, Feb 2012
36. A. Mizuno, K. Kohno, R. Ohyama, T. Tokuyoshi, H. Uetani, H. Eichel, T. Miyamori, N. Matsumoto, M. Matsui, Design methodology and system for a configurable media embedded processor extensible to VLIW architecture, in *International Conference on Computer Design*, pp. 2–7, Sept 2002
37. OpenCores: miniMIPS, http://opencores.org/project,minimips. Accessed 16 Sept 2013
38. H. Onodera, A. Hirata, T. Kitamura, K. Tamaru, P2lib: process-portable library and its generation system, in *Custom Integrated Circuits Conference*, pp. 341–44, May 1997
39. T. Sato, T. Hayashida, K. Yano, Dynamically reducing overestimated design margin of multicores, in *10th International Conference on High Performance Computing & Simulation*, pp. 403–409, July 2012
40. M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, T. Kawahara, 90-nm process-variation adaptive embedded SRAM modules with power-line-floating write technique. IEEE J. Solid-State Circ. **41**(3), 705–711 (2006)
41. S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, H. Shinohara, A 65-nm soc embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits. IEEE J. Solid-State Circ. **42**(4), 820–829 (2007)
42. M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi, K. Hashimoto, A. Sebe, G. Okazaki, K. Satomi, H. Akamatsu, H. Shinohara, A 45 nm low-standby-power embedded SRAM with improved immunity against process and temperature variations, in *IEEE ISSCC Digest of Technical Papers*, pp. 326–327, Feb 2007

43. K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, Y. Oda, K. Usui, T. Kawamura, N. Tsuboi, T. Iwasaki, K. Hashimoto, H. Makino, and H. Shinohara, A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment, in *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 212–213, June 2008
44. Y. Fujimura, O. Hirabayashi, T. Sasaki, A. Suzuki, A. Kawasumi, Y. Takeyama, K. Kushida, G. Fukano, A. Katayama, Y. Niki, T. Yabe, A configurable SRAM with constant-negative-level write buffer for low-voltage operation with 0.149 um^2 cell in 32 nm high-k metal gate CMOS, in *IEEE ISSCC Digest of Technical Papers*, pp. 348–349, Feb 2010
45. T. Yabe et al., Circuit techniques to improve disturb and write margin degraded by MOSFET variability in high-density SRAM cells, in *Symposium on VLSI Circuits Digest of Technical Papers*, June 2011
46. H. Pilo, I. Arsovsi, K. Batson, G. Braceras, J. Gabric, R. Houle, S. Lamphier, C. Radens, A. Seferagic, A 64 Mb SRAM in 32 nm high-k metal-gate SOI technology with 0.7 V operation enabled by stability, write-ability and read-ability enhancements. IEEE J. Solid-State Circ. **47** (1), Jan 2012
47. E. Seevinck, F.J. List, J. Lohstroh, Static-noise margin analysis of MOS SRAM cells. IEEE J. Solid-State Circ. **SC-22**(5), 748–754 (1987)
48. R. Heald, P. Wang, Variability in sub-100 nm SRAM designs, in *IEEE/ACM ICCAD Digest of Technical Papers*, pp. 347–352, Nov 2004
49. M. Khellah, Y. Ye, N. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, V. De, Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs, in *Symposium VLSI Circuits Digest Technical Papers*, pp. 9–10, June 2006
50. H. Pilo, C. Barwin, G. Braceras, C. Browning, S. Lamphier, F. Towler, An SRAM design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage. IEEE J. Solid-State Circ. **42**(4), 813–819 (2007)
51. M. Yamaoka, K. Osada, T. Kawahara, A cell-activation-time controlled SRAM for low-voltage operation in DVFS SoCs using dynamic stability analysis, in *Proceedings of European Solid-State Circuits Conference* (*ESSCIRC*), pp. 286–289, Sep 2008
52. H. Nho, P. Kolar, F. Hamzaoglu, Y. Wang, E. Karl, Y. Ng, U. Bhattacharya, K. Zhang, A 32 nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation, in *IEEE ISSCC Digest of Technical Papers*, pp. 346–347, Feb 2010
53. E. Karl, Y. Wang, Y.-G. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, M. Bohr, A 4.6 GHz 162 Mb SRAM design in 22 nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry, in *IEEE ISSCC Digest of Technical Papers*, pp. 230–231, Feb 2012
54. H. Nakano, T. Iwao, T. Hishida, H. Shimomura, T. Izumi, T. Fujino, Y. Okuno, K. Arimoto, An embedded programmable logic matrix (ePLX) for flexible functions on SoC, in *IEEE ASSCC Digest of Technical Papers*, pp. 219–222, Nov 2006
55. K. Nii, M. Yabuuchi, H. Fujiwara, H. Nakano, K. Ishihara, H. Kawai, K. Arimoto, Dependable SRAM with enhanced read/write-margins by fine-grained assist bias control for low-voltage operation, in *Proceedings of IEEE International SOC Conference*, pp. 519–524, Sept 2010