Low Power Testing

Patrick Girard, Xiaoqing Wen, Nur Touba

Table of Contents

7. Low-Power Testing
– by Patrick Girard, Xiaoqing Wen, and Nur A. Touba

Chapter

7

Low-Power Testing

Patrick Girard, *LIRMM/CNRS, Montpellier, France*
Xiaoqing Wen, *Kyushu Institute of Technology, Fukuoka, Japan*
Nur A. Touba, *University of Texas, Austin, Texas, USA*

About This Chapter

Power dissipation has become a major design objective in many application areas, such as wireless communications and high performance computing, thus leading to the production of numerous low-power designs. At the same time, power dissipation is also becoming a critical parameter during manufacturing test, as the design can consume much more power during test than during functional mode of operation. Because test throughput and manufacturing yield are often affected by **test power**, dedicated test methodologies have emerged over the past decade.

In this chapter, we discuss issues arising from excessive power consumption during test application as well as provide structural and algorithmic solutions that can be used to alleviate the low-power test problems. We first review some basic elements of power modeling and related terminologies. After discussing test power issues, promising low-power test techniques to deal with nanometer ***system-on-chip*** (SOC) designs are presented. These techniques can be broadly classified into those that apply during **scan** testing and those that apply during ***built-in self-test*** (BIST). A few of them are also applicable to test compression circuits or memory designs.

In the literature, techniques that reduce power consumption during test application are generally referred to as *power-conscious testing*, *power-aware testing*, *power-constrained testing*, or *low-power testing*. These terms will be interchanged for use throughout the chapter whenever fit.

7.1 Introduction

With the advance in semiconductor manufacturing technology, a ***very-large-scale-integration*** (VLSI) device can now contain tens to hundreds of millions of transistors. Because this trend is predicted to continue at least for the next 10 years per Moore's law [Moore 1965], severe challenges are imposed on tools and methodologies used to design and test complex VLSI circuits. Addressing these design and test challenges in an efficient way is now becoming increasingly difficult [SIA 2005].

Test currently ranks among the most important issues in the development process of an integrated circuit. The issues that center on test are manufacturing yield, product quality, and test cost. To address these test issues, ***design-for-testability*** (DFT) techniques [Bushnell 2000] [Jha 2003] [Wang 2006] have become widely used in industry since the 1990s. Traditionally, these techniques are mainly employed to improve the circuit's fault coverage, test application time, and test development efforts. The recent advances in low-power design techniques and deep-submicron manufacturing technologies, however, have spurred the rapid growth of electronic products into consumer markets using laptop computers, cellular phones, audio and video-based multimedia products, energy-efficient desktop computers, etc. These new products make power management a critical issue that needs to be considered not only during circuit design but also during test development [Crouch 1999] [De Colle 2005].

The main motivation for considering power consumption during test is that generally, a circuit consumes much more power in test mode than in normal mode [Zorian 1993] [Rajski 1998] [Girard 2000] [Pouya 2000] [Bushnell 2000] [SIA 2001] [Saxena 2003] [Nicolici 2003]. It was shown in [Zorian 1993] that test power can be more than twice the power consumed in normal functional mode. There are several reasons that could explain this increase in **test power**. First, modern automatic test pattern generation (ATPG) tools tend to generate test patterns with a high toggle rate in order to reduce pattern count and thus test application time. Thus, the node switching activity of the device in test mode is often several times higher than that in normal mode. Second, parallel testing (e.g., testing a few memories in parallel) is often used to reduce test application time, particularly for ***system-on-chip*** (SOC) devices. This parallelism inevitably increases power dissipation during test. Third, the DFT circuitry inserted in the circuit to alleviate test issues is often idle during normal operation but may be intensively used in test mode. This surplus of active elements during test again induces an increase of power dissipation. Finally, this elevated test power can come from the lack of correlation between consecutive test patterns, while the correlation between successive functional input vectors applied to a given circuit during normal operation is generally very high [Wang 1997].

For instance, in a speech signal processing circuit, the input vectors behave in a predictable manner, with the least significant bits more likely to change than the most significant bits. Similarly, in high-speed circuits that process digital audio and video signals, the inputs to most of those modules change relatively slowly over time. The low-power designers often take advantage of this fact when they determine the thermal and electrical limits of the circuit and system packaging requirements. In contrast, there is no definite correlation between successive test patterns generated by an ATPG tool during scan testing or produced by a ***pseudo-random pattern generator*** (PRPG) during logic BIST. As power dissipation in CMOS circuits is proportional to switching activity, this excessive switching activity during test can cause catastrophic problems, such as instant circuit damage, test-induced yield loss due to noise phenomena, reduced reliability, product cost increase, or reduced autonomy for battery-operated devices.

In order to reduce this increased power consumption during test application, the industry generally resorts to *ad hoc* solutions [Monzel 1997]. These solutions include:

- ♦ Over-sizing power and ground rails to allow higher current densities in the circuit under test. This allows additional power to be supplied to the circuit to satisfy the increase in switching activity that occurs during test. However, this solution raises several problems. By increasing the power available for the circuit, the amount of energy (heat) that needs to be dissipated is also increased, which in turn leads to additional problems related to the thermal constraints of the circuit (these problems are discussed in Section 7.3). It is possible to avoid these problems by using packages with higher thermal capabilities or by using higher performance cooling systems. However, the impact on the final product cost may prevent the use of these solutions. Another problem is that this solution affects the entire design and may require an early estimation of the power consumption during test. As test data is generally not available in the early stages of the design process, this solution may not be satisfactory in all cases.

- ♦ Testing with a reduced operating frequency. This solution does not require additional hardware, but it increases the test application time and may lead to a loss of defect coverage as timing-related faults may escape detection. In effect, this solution reduces power consumption at the expense of longer test time, and does not reduce the total energy consumed during test.

- ♦ Partitioning of the circuit under test with appropriate test planning. This solution, although effective from a power reduction point of view, increases test time because it reduces test concurrency. Moreover, it generally requires circuit design modifications (often with additional multiplexers), thus impacting final product cost and circuit performance.

Considering the problems associated with these *ad hoc* approaches and the need to provide an adequate remedy to the problems, numerous solutions have been proposed in recent years to cope with test power problems during test. These solutions can be classified based on whether they apply during scan testing or whether they apply during logic BIST. A few of them can also be used with memory designs or can be used in conjunction with test compression. These solutions are explained in detail in the next sections.

7.2 Energy and Power Modeling

A logical step before discussing promising low-power test solutions is to correctly define the terminology and the associated energy and power models. In this section we first review the electronic basics related to power consumption and power dissipation and then proceed to the discussion of terminology and test power modeling.

7.2.1 Basics of Circuit Theory

Consider the generic representation of a ***complementary metal-oxide semiconductor*** (CMOS) logic gate shown in Figure 7.1. The load capacitance C_L of the output node, representing the input capacitance of the next logic stage as well as interconnect and diffusion capacitances, is connected to the supply voltage V_{dd} through a pull-up block composed of ***positive metal-oxide semiconductor*** (PMOS) transistors and to the ground through a pull-down block composed of ***negative metal-oxide semiconductor*** (NMOS) transistors.

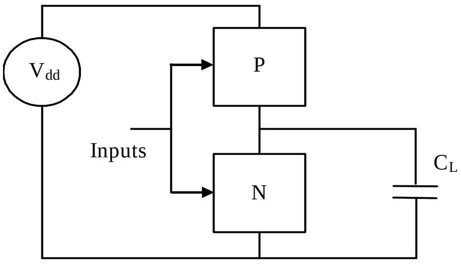


Figure 7.1: Generic representation of a CMOS logic gate

A switching on the gate output corresponds to the charge or discharge of the load capacitance C_L. In the process of charging the output (from 0 to 1), a charge $Q = C_L.V_{dd}$ is delivered to the load. The power supply must supply this charge at voltage V_{dd}, so the energy supplied is $Q.V_{dd} = C_L.V_{dd}^2$. However, the energy stored on a capacitance C_L charged to V_{dd} is only half of this, i.e., $½.C_L.V_{dd}^2$ [Athas 1994]. In accordance with the **energy conservation principle**, the other half must be dissipated by the PMOS transistors in the pull-up network. Similarly, when the inputs change again causing the output to discharge (from 1 to 0), all the energy stored on the capacitance C_L is inevitably dissipated in the pull-down network, because no energy can enter the ground rail ($Q.V_{gnd} = Q.0 = 0$). In both cases, the energy is dissipated as heat.

There are three components to the power consumed by the logic gate: (1) the **dynamic power**, due to the charge of capacitance C_L, (2) the **short-circuit power**, due to the short circuit between power and ground during switching, and (3) the **leakage power**. The main component is the dynamic power, which still represents a significant fraction of the total power consumption despite the proportional increase of the other two components with technology improvements. This dynamic power consumption occurs during the charge of the load capacitance C_L (transition from 0 to 1 on the gate output) as a current I flows between power and ground through the capacitance. The dynamic power consumed during the time interval $[0,T]$ is therefore: $P_{dyn} = V_{dd} \bullet I = V_{dd} \bullet Q \bullet 1/T$ where $Q = C_L \bullet V_{dd}$. As several transitions may occur during the time interval $[0,T]$, the dynamic power consumption can be expressed as follows:

$$P_{dyn} = C_L.V_{dd}^2.N_{0\rightarrow 1}.1/T$$

where $N_{0\rightarrow 1}$ represents the number of rising transitions at the gate output during the time interval $[0,T]$. Without loss of generality, it can be assumed that the number of rising transitions is equal to half of the total number of N transitions at the gate output. The dynamic power consumption of the logic gate during the time interval $[0,T]$ can finally be expressed as:

$$P_{dyn} = \tfrac{1}{2}.C_L.V_{dd}^2.N.1/T$$

The above analysis shows that **dynamic power consumption** occurs during the charge of node output capacitance, whereas **power dissipation**, which is related to **energy dissipation**, occurs during the charge or discharge of each node. Because power dissipated by N rising or falling transitions during the time interval $[0,T]$ is $E_n/T = \tfrac{1}{2} \bullet C_L \bullet V_{dd}^2 \bullet N \bullet 1/T$, which is the same as power consumption, the terms power dissipation and power consumption will be used without distinction throughout this chapter.

7.2.2 Terminology

We use the same terminology as defined in [West 1993] to denote power consumption measures used for low-power testing:

- **Energy** represents the total switching activity generated during the application of the complete test sequence. An energy increase during test has impact on the battery lifetime of battery operated devices, particularly those equipped with on-line test facilities or those submitted to test procedures during power up (such as cellular phones).

- **Average Power** corresponds to the ratio between the total energy dissipated during test and the test time. Elevated average power during test adds to the thermal load that must be vented away from the device under test (temperature increase). It may cause structural damage to the silicon (hot spots) or leads to phenomena that alter the circuit reliability.

- **Instantaneous Power** corresponds to the power consumed at any given instant during test. Usually, it is defined as the power consumed right after the application of a synchronizing clock signal. Elevated instantaneous power may cause a supply voltage drop and alter the correct behavior of the circuit.

- **Peak Power** corresponds to the highest value of instantaneous power measured during test. The peak power generally determines the thermal and electrical limits of the circuit and the system packaging requirements. If the peak power exceeds a certain limit, the circuit may be subjected to structural degradation and, in some cases, be destroyed. From a theoretical point of view, the peak power is defined from the values of instantaneous power measured on very short time intervals, i.e., the system clock period. In practice, the time window for the definition of peak power is related to the thermal capacity of the chip, and restricting this window within just one clock period is not realistic enough. For example, if the circuit has a peak power consumption during only one cycle but it has power consumption within the limit of thermal capacity of the chip for all other cycles, the circuit may not be damaged because the energy consumed may not be enough to elevate chip temperature over the limit of thermal capacity of the chip (unless the peak power consumption is far higher than normal power consumption). To damage the circuit, high power consumption should last for several successive clock cycles to consume enough energy to elevate chip temperature over the limit [Shi 2004]. On the other hand, high peak power in only one clock cycle can be an issue if it results in a significant ground bounce or an IR-drop phenomenon that causes a memory element to lose its state and the test procedure to unnecessarily fail. This problem will be further discussed in Section 7.3.2.

7.2.3 Test-Power Modeling and Evaluation

As mentioned above, most power dissipated in a CMOS circuit comes from the charge and discharge of capacitances during switching. In order to explain this power dissipation during test, let us consider a circuit composed of N nodes, and a test sequence of length L applied to the circuit inputs. The **average energy** consumed at node i per switching is $½.C_i.V_{dd}^2$ where C_i is the equivalent output capacitance at node i and V_{dd} the power supply voltage [Cirit 1987]. A good approximation of the energy consumed at node i in a time interval t is $½.C_i.S_i.V_{dd}^2$ where S_i is the average number of transitions during this interval (also called the **switching activity factor** at node i). Furthermore, nodes connected to more than one logic gate in the circuit are nodes with a higher output capacitance. Based on this fact, and in a first approximation, it can be stated that output capacitance C_i

is proportional to the fanout at node *i*, denoted as F_i [Wang 1995]. Therefore, an estimation of the energy E_i consumed at node *i* during the time interval *t* is given by:

$$E_i = \tfrac{1}{2}.S_i.F_i.C_0.V_{dd}^2$$

where C_0 is the minimum output capacitance of the circuit. According to this expression, energy consumption at the logic level is a function of the fanout F_i and the switching activity factor S_i. The fanout F_i is defined by circuit topology, and the activity factor S_i can be estimated by a logic simulator. The product $F_i.S_i$ is named ***weighted switching activity*** (WSA) at node *i* and represents the only variable part in the energy consumed at node *i* during test application.

According to the above formulation, the energy consumed in the circuit after application of a pair of successive input vectors (V_{k-1},V_k) can be expressed by:

$$E_{Vk} = \tfrac{1}{2}.C_0.V_{dd}^2.\sum_i S_i(k).F_i$$

where *i* ranges across all the nodes of the circuit and $S_i(k)$ is the number of transitions provoked by V_k at node *i*. Now, let us consider the complete test sequence of length *L* required to achieve the target fault coverage. The **total energy** consumed in the circuit after the application of the complete test sequence is given below, where *k* ranges across all the vectors of the test sequence.

$$E_{total} = \tfrac{1}{2}.C_0.V_{dd}^2.\sum_k \sum_i S_i(k).F_i$$

By definition, power is given by the ratio between energy and time. The instantaneous power is generally calculated as the amount of power required during a small instant of time t_{small} such as the portion of a clock cycle immediately following the system clock rising or falling edge. Consequently, the instantaneous power dissipated in the circuit after the application of a test vector V_k can be expressed by:

$$P_{inst}(V_k) = E_{Vk} / t_{small}$$

The peak power corresponds to the maximum value of instantaneous power measured during test. Therefore, it can be expressed in terms of the highest energy consumed during a small instant of time during the test session:

$$P_{peak} = Max_k P_{inst}(V_k) = Max_k(E_{Vk} / t_{small})$$

Finally, the average power consumed during the test session can be calculated from the total energy and the test time. Considering that the test time is given by the product *L.T*, where *T* corresponds to the nominal clock period of the circuit, the average power can be expressed as follows:

$$P_{average} = E_{total} / (L.T)$$

The above expressions of power and energy, although based on a simplified model, are accurate enough for the intended purpose of power analysis during test. According to these expressions, and assuming a given CMOS technology and a supply voltage for the considered circuit, it appears that the switching activity factor S_i is the only parameter that has impact on the energy, peak power, and average power. This explains why most of the methods proposed so far for reducing power and/or energy during test are based on a reduction of the switching activity factor.

7.3 Test Power Issues

When verifying the correct functions of high-density systems such as an SOC, test procedures and test techniques have to satisfy all power constraints defined in the design phase. In other words, these procedures and techniques must be so that the power consumed during test remains comparable to that consumed during functional mode. Ignoring these constraints can expose the circuit to various problems, such as premature destruction, noise phenomena that can lead to yield loss, reduced reliability, product cost increase, reduced autonomy (for battery-operated devices), etc. This section lists a few of these important problems.

7.3.1 Thermal Effects

The heat produced during the operation of a circuit is proportional to the dissipated power. This heat is produced by the collision of carriers with the conductor molecular structure (a friction phenomenon called the **Joule effect**), and is responsible for the temperature increase observed during operation [Altet 2002]. Therefore, there is a relationship between die temperature and power dissipation. It can be formulated from the **Laws of Thermodynamics** as follows [West 1993]:

$$T_{die} = T_{air} + \theta \times P_d$$

where T_{die} is the *die temperature*, T_{air} is the temperature of surrounding air, θ is the *package thermal impedance* expressed in °C/Watt, and P_d is the *average power* dissipated by the circuit. From this expression, it is clear that an excessive power dissipated during test will increase the circuit temperature well beyond the value measured (or calculated) during functional mode [SIA 2003]. If the temperature is too high, even during the short duration of a test session, it can result in irreversible structural degradations. Some of these degradations, such as **hot spots**, appear during test data application and may lead to premature destruction of the circuit [Pouya 2000]. Some others, which are accelerated gradually over time (*ageing*), may affect circuit performance or cause functional failures after a given lifetime [Hertwig 1998] [Shi 2004]. In this case, the main mechanisms leading to these structural degradations are *corrosion* (oxidizing of conductors), *electromigration* (molecular migration of the conductor structure towards the electronic flow), *hot-carrier-induced* defects, or *dielectric breakdown* (loss of insulation of the dielectric barrier) [Altet 2002]. These types of degradations have a big impact on long-term circuit reliability.

7.3.2 Noise Phenomena

These types of problems can occur when testing the circuit at the wafer level (for *characterization testing* or *verification testing*). For this type of test, the power must be supplied to the circuit through probes which typically have higher inductances than the power and ground pins of the package planned for circuit encapsulation. If the switching activity during test is equal to or higher than the switching activity during functional mode, the **power supply noise** (which is given by $L(di/dt)$ where L is the inductance of a power line and di/dt represents the magnitude of the variation of the current flowing through this line) will be increased [Wang 1997]. This excessive noise can erroneously change the logic state of some circuit nodes at a given instant, causing some good dies to fail the test, thus leading to unnecessary loss of yield. In order to avoid such phenomena, it is important to reduce test power.

Comparable inductive phenomena, known as "**ground bounce**" or "**voltage surge/droop**", may occur during testing of the packaged circuit (*production testing*). Actually, wire/substrate inductances or package lead inductances associated with power or ground rails appear in circuits designed with deep submicron technologies. When very high switching currents occur in the circuit under test, caused by high switching activity, voltage glitches can be observed at the nodes of these inductances [Jiang 2000]. These voltage glitches are proportional to both the inductance value and the magnitude of the variation of the current flowing through this inductance. In some cases, these voltage glitches may change the rise/fall times of some signals in the circuit (timing performance degradation). In other cases, they can erroneously change the logic state of some circuit nodes or flip-flops, and cause some good dies to fail the test thus leading to yield loss [Chang 1997]. Once again, high switching rates, elevated operating frequencies, and short rise/fall times of internal signals are the primary cause of these phenomena, worsened by the increased susceptibility of today's circuits to these noise phenomena.

Similarly, **IR-drop** and crosstalk effects are noise phenomena that may show up as an error in test mode but not in functional mode. IR-drop refers to the amount of decrease (increase) in the power (ground) rail voltage and is linked to the existence of a non negligible resistance between the rail and each node in the circuit under test. **Crosstalk** refers to capacitive coupling between neighboring nets within an IC. With high peak current demands during test, the voltages at some gates in the circuit are reduced. This causes these gates to exhibit higher delays, possibly leading to test fails and yield loss [Butler 2004]. These phenomena have been widely reported in the literature, in particular when at-speed transition delay testing is performed [Shi 2004]. Typical examples of voltage drop sensitive applications are Gigabit switches containing millions of logic gates.

7.3.3 Miscellaneous Issues

The **cost** constraints of consumer electronic products typically require the use of plastic packages for integrated circuit packaging. This type of package, although quite cheap, is not always able to dissipate high levels of heat. However, the use of packages with higher thermal capacities, such as ceramic or organic packages, which would allow removal of the excessive heat dissipated during test, would significantly increase the final product cost. Similarly, the use of special cooling systems that could be used for venting away the excess of heat generated during test, such as a radiator or fan, would also have a negative impact on the product cost. Moreover, in portable systems, where weight and size are very important, these solutions are completely out of the question. Thus, it is important to reduce test power in order to avoid cost increases in these types of products.

Embedded electronic systems powered by batteries are employed in various types of applications (computing, aerospace, avionics, telephony, automotive, military, etc.). In mission-critical and safety-critical applications (e.g., avionics and aerospace), these systems are equipped with BIST features to periodically check that the circuits are functioning correctly by taking advantage of idle periods in the system operation [Nicolaidis 1998]. For applications such as telephony, power-up self-test procedures are used to check the system integrity and alert the user when problems occur. In this case, test resources are also embedded in the system to facilitate such operations. For all these applications, **autonomy** is a critical issue that needs to be addressed during test by minimizing the switching activity. As the main issue here is the amount of energy used, it is also possible to minimize the impact of test by reducing the length of the test sequences used.

Finally, another reason why it is important to reduce power consumption during test is the need for applying at-speed tests. In the past, tests were typically applied at rates much lower than the functional clock rate of the circuit. The main goal was to screen static faults (such as stuck-at faults). Thus, the excess of switching activity generated during test was compensated by the reduction of the test clock frequency. Today, timing defects are becoming prevalent due to the use of nanometer process technology. This makes it essential to test for delay faults to ascertain circuit performance. Therefore, tests have to be applied at-speed, and it is no longer practical to reduce the test clock frequency [Krstic 1998]. Minimizing switching activity during test for reducing power consumption thus becomes imperative.

7.4 Low-Power Scan Testing

In the context of scan testing, the problem of excessive power during test is much more severe than in functional mode. This is mostly due to the fact that the application of each test pattern in a scan design requires a number of shift clock cycles that contributes to an unnecessary increase of switching activity [Bushnell 2000] [Wang 2006]. A study reported in [Saxena 2003] shows that while 10%-20% of the memory elements (D flip-flops and D latches) in a digital circuit change state during one clock cycle in functional mode, 35%-40% of these memory elements when reconfigured as scan cells can switch state during scan testing. In the worst case, all scan cells can switch state. Another report [Shi 2004] further indicates that the average power during scan testing can be 3 times the power consumed during normal functional operation, and the peak power can be 30 times what it is in normal functional operation. In this section, we discuss various low-power scan test techniques to reduce excessive test power.

7.4.1 Basics of Scan Testing

Scan design requires reconfiguration of memory elements (often D flip-flops) into scan cells and then stitching them together to form scan chains [Bushnell 2000] [Jha 2003] [Wang 2006]. During **slow-speed scan testing**, each scan test pattern must be first shifted into the scan chains. This requires setting the scan cells to shift mode and applying a number of load/unload (shift) clock cycles. Scan shifting is generally done at slow speed in order to avoid high power dissipation. A capture clock cycle is then applied to capture the test response of the design into scan cells. This requires setting the scan cells to normal/capture mode. A scan enable signal (*SE*) is typically used for this setting. When *SE* is set to 1, the scan design is in shift mode; when *SE* is set to 0, the circuit is switched to normal/capture mode.

Since the early 1990s, this slow-speed scan test technique has been widely used in industry to test stuck-at faults and bridging faults. The shift and capture operations for a typical scan design with the associated current waveform for each clock cycle are shown in Figure 7.2. This current waveform varies cycle by cycle because current is proportional to the number of 0-to-1 and 1-to-0 transitions on the scan cells which in turn produce switching in the ***circuit under test*** (CUT).

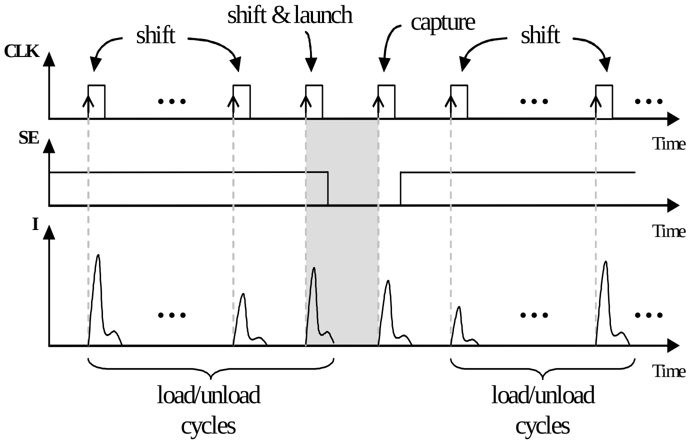


Figure 7.2: Slow-speed scan testing with associated current waveform

The problem of excessive power during (slow-speed) scan testing can be split into two sub-problems: excessive power during the shift operation (called **excessive shift power**) and excessive power during the capture operation (called **excessive capture power**) [Girard 2002]. The latter is intended to address clock skew problems that may arise when many flip-flops change their output values simultaneously after capture operation. Over the past ten years, numerous techniques have been proposed to reduce shift power, capture power, or both at the same time during slow-speed scan testing. These low-power scan test techniques are introduced in detail in the following subsections.

In the meantime, as feature size shrinks into the ***deep-submicron*** (DSM) scale and circuit speed starts to operate at the *GHz* range, we have seen more and more chips fail due to timing-related defects. As a result, **at-speed scan testing**, which captures test response of the scan design at the rated clock speed, is becoming mandatory to ensure high product quality.

There are two types of at-speed scan test schemes: ***launch-on-shift*** (LOS) and ***launch-on-capture*** (LOC) [Wang 2006]. Figure 7.3 shows the clock diagram of both schemes. Using the launch-on-shift scheme, vector V_1 (after the next to last shift) is loaded to the scan cells for initialization and a second vector V_2 (after the last shift) is then shifted into the scan cells to launch a transition on selected scan cells in shift mode. In this case, V_2 is a one-bit shift of the first vector V_1. One capture clock cycle is then applied at-speed to the design in normal/capture mode to capture the test response. On the other hand, using the launch-on-capture scheme, two capture clock cycles (launch and capture pulses as shown in Figure 7.3) are applied at speed in normal/capture mode to capture the final test response to scan cells. In this case, V_2 is the circuit's response to V_1 which is then captured to the scan cells in capture mode again. Experiments have shown that a LOS test can have higher delay fault coverage than a LOC test [Xu 2006]. However, since

LOS requires an at-speed scan enable signal (*SE*), it is more difficult to layout the scan design. Moreover, LOS suffers from an overkill issue which could reject good chips as more false paths can be activated than using the LOC scheme.

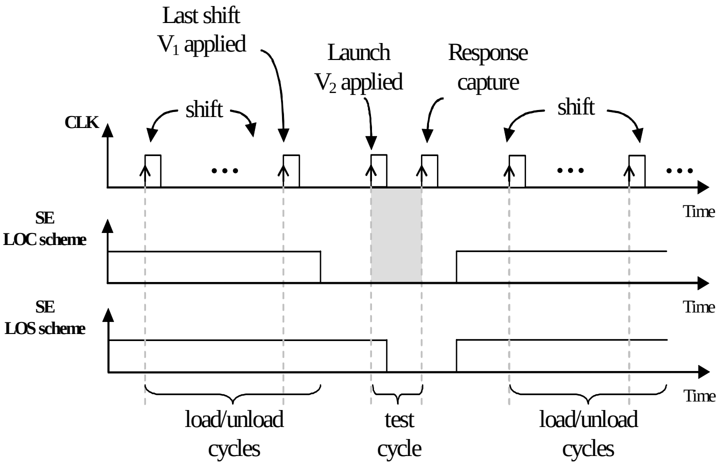


Figure 7.3: At-speed scan testing with LOS and LOC test schemes

The applicability of at-speed scan testing is also further challenged by **test-induced yield loss**, an emerging test problem which occurs when good chips fail only during test. The main source of test-induced yield loss is **excessive IR-drop** caused by the high switching activity generated in the CUT between launch of the test stimulus and capture of the corresponding test response [Butler 2004]. Excessive IR-drop, during the short period between launch and capture (called the **test cycle**), may lead to a situation where gates in the circuit exhibit higher delays so an erroneous response may be captured to the scan cells at the end of the test cycle (the response capture edge). This makes at-speed scan testing especially vulnerable to IR-drop. A few solutions, based on power-aware ATPG or X-filling and presented in Section 7.4.2, have been proposed to avoid IR-drop-induced yield loss.

7.4.2 ATPG and X-Filling Techniques

In conventional scan ATPG, each don’t care bit (*X*) in a **test cube** is filled with 0 or 1 randomly; the resulting fully-specified test cube (called a scan test pattern or simply test pattern) is then fault-graded to confirm the detection of all targeted faults and additional faults. While state-of-the-art dynamic and static test pattern compaction techniques have been extensively used to reduce pattern count in scan ATPG, the number of don’t care bits in a given test cube remains high [Wohl 2003] [Hiraide 2003]. This provides a great opportunity that can be exploited for power minimization during scan testing.

The first set of techniques to reduce test power is to use novel low-power ATPG algorithms for generating low-power test patterns that still meet the original ATPG objectives (maximum fault coverage and minimum pattern length with reasonable run time). The authors in [Wang 1994] enhanced the ***path-oriented decision-making*** (PODEM) algorithm by assigning don't care bits present at the CUT inputs in a clever manner to minimize the number of transitions between two consecutive test patterns. This reduces both average and peak power dissipation during shift operations. In [Wang 1997], the authors further extend the ATPG approach proposed in [Wang 1994] to full-scan sequential circuits by exploiting all don't cares that occur during scan shifting, test application, and response capture to minimize shift power and capture power simultaneously.

Another low-power ATPG method for efficient capture power reduction during scan testing [Wen 2006] tries to achieve two goals: the primary one being the detection of targeted faults and the secondary one being the minimization of the difference between before-capture and after-capture output values of scan cells. This is achieved by introducing the concept of a capture conflict (*C-conflict*) in addition to the conventional detection conflict (*D-conflict*). A C-conflict occurs when a difference between the before-capture and after-capture output values of a scan cell is created by logic value assignment during ATPG. A C-conflict, in the same manner as a D-conflict, may be avoided through the backtrack operation. However, backtracking for a C-conflict may make fault detection impossible. In this case, the backtracking for the C-conflict is reversed, and the transition at the scan cell is tolerated since the primary goal is fault detection.

The second set of techniques to reduce test power is to use **power-aware *X*-filling** heuristics that do not modify the overall ATPG process. Given a set of deterministic test cubes, the main goal of these techniques is to assign values to the don't care bits of each test cube so that the number of transitions in the scan cells is minimized. By reducing the number of transitions in the scan cells during scan shifting, the overall switching activity in the CUT is also reduced; power consumption during test is thus minimized. Most of the time, the *X*'s are assigned with the help of the following classical *non-random filling heuristics*:

- ***Minimum transition filling*** (MT-filling), also call **Adjacent filling**: all don't care bits in a test cube are set to the value of the last encountered care bit. That is, when applying MT-filling, the most recent care bit value is used to fill successive *X* values until a care bit is reached.
- **0-filling**: all don't care bits in a pattern are set to '0'.
- **1-filling**: all don't care bits in a pattern are set to '1'.

MT-filling results in the fewest number of transitions in the scan chains which generally corresponds to the lowest switching activity in the overall circuit, and is thus the preferred approach. Consider the test cube <0*XXX*1*XX*0*XX*0*XX*>. By applying the above three non-random filling heuristics, the resulting patterns become:

- 0000111000000 with MT-filling heuristics
- 0000100000000 with 0-filling heuristics
- 0111111011011 with 1-filling heuristics

These classical non-random filling heuristics (among a few others) have been evaluated [Butler 2004] to measure the reduction in average power consumption during scan shifting (load/unload cycles). These heuristics have also been evaluated to measure the reduction in peak power consumption with respect to a random filling of don’t care bits [Badereddine 2006]. Complete results on benchmark circuits have shown that both average and peak power consumption during test can be efficiently minimized with the MT-filling heuristics.

In the context of at-speed scan testing, a few *X*-filling solutions have also been described to reduce power during the test cycle and thus avoid IR-drop-induced yield loss [Wen 2005a] [Wen 2005b] [Remersaro 2006]. These solutions have been developed to provide power-aware LOC delay tests. The basic idea is to minimize the bit differences between V_1, the initialization vector, and V_2, the sensitizing vector (which in this case is equal to the output response of V_1), while maintaining the original transition fault coverage.

Compared to other solutions, X-filling techniques have the advantage of being applicable at the end of the design process (without imposing any impact on the design flow) and thus do not require any modification of the circuit and hence do not incur any area overhead. These methods reduce test power consumption sometimes at the expense of an increase in the pattern count due to the fact that they may not be as effective in detecting additional faults as random filling thereby requiring incrementally more patterns to achieve the target fault coverage.

7.4.3 Low-Power Test Vector Compaction

Static compaction involves minimizing the number of test cubes generated by an ATPG tool by merging test cubes that are compatible in all bit positions (i.e., they have no conflicting bit position where one test cube has a specified ‘1’ and another has a specified ‘0’). Conventional approaches for static compaction merge test cubes in an arbitrary order until no more merging is possible. It was shown in [Sankaralingam 2000] that by carefully selecting the order in which test cubes are merged, the number of transitions can be minimized. To measure the number of transitions that result when shifting a scan vector into a scan chain, a **weighted transition metric** can be used. To illustrate the weighted transition metric, consider the example given in Figure 7.4. It has two transitions. When this vector is scanned into the CUT, Transition 1 passes through the entire scan chain. This transition dissipates power at every scan cell in the scan chain. On the other hand, Transition 2 dissipates power only at the first scan cell during scan in. The number of scan cell transitions caused by a transition in a scan vector being scanned in depends on its position in the scan vector. In this example where there are 5 scan cells, a transition in position 1 (which is where Transition 1 is) would be weighted 4 times more than a transition in position 4 (which is where Transition 2 is). The weight assigned to a transition is the difference between the size of the scan chain and the position in the vector in which the transition occurs. Hence, the power dissipated when applying two vectors can be compared by counting the number of weighted transitions in each vector. The number of weighted transitions is given by:

$$Weighted\_Transitions = \Sigma(Size\_of\_Scan\_Chain - Position\_of\_Transition)$$

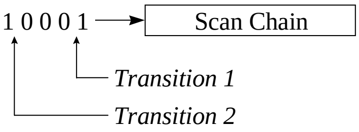


Figure 7.4: Transitions in scan vector

Using this metric, a greedy heuristic procedure is given in [Sankaralingam 2000] for merging test cubes in a way that minimizes the number of transitions. Significant reductions in average and peak power consumption can be obtained by using this approach.

7.4.4 Shift Control Techniques

Several techniques have also been proposed to reduce or cancel the switching activity in the CUT during scan shifting. The authors in [Huang 1999] try to find an input vector, called a *control vector*, such that when this vector is applied to the primary inputs of the CUT during scan shifting, the switching activity in the combinational part of the CUT is minimized. To determine this input control vector, a modified version of the D-Algorithm is used. The method has achieved some reasonable reduction in average power consumption.

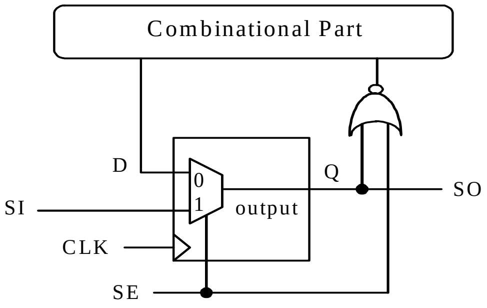


Figure 7.5: Scan cell modification

Another technique proposed in [Hertwig 1998] is to modify each scan cell in the scan chain so as to block transitions at the scan cell outputs during scan shifting and thereby prevent all switching activity in the combinational portion of the

CUT. The scan cell modification consists of adding a NOR gate and an additional fanout to the output of each scan cell (see Figure 7.5). During scan shifting, the NOR gate prevents data in the scan cell from propagating to the combinational part of the CUT. This technique obviously is very effective in test power reduction; however, it requires a significant area overhead and may degrade the circuit performance.

7.4.5 Scan Cell Ordering

In scan design, switching activity, and hence power dissipation, can be further reduced by changing the order of the scan cells in each scan chain. Consider a scan chain composed of four ordered scan cells (FF_1-FF_2-FF_3-FF_4). Assume that the test vector V=<0101> is to be loaded in the scan chain and the initial state of the four scan cells is <0000> in this scan chain. The total number of transitions generated in the scan chain by the loading of vector V will be equal to 10 (see Figure 7.6a). Now, suppose that the order of the scan cells in the scan chain is changed to FF_2-FF_4-$F\mathrm{F}_1$-FF_3, the total number of transitions in the scan chain in this case becomes 2 (see Figure 7.6b). Of course, changing the order of the scan cells in the scan chain implies a change of the bit order in each test vector.

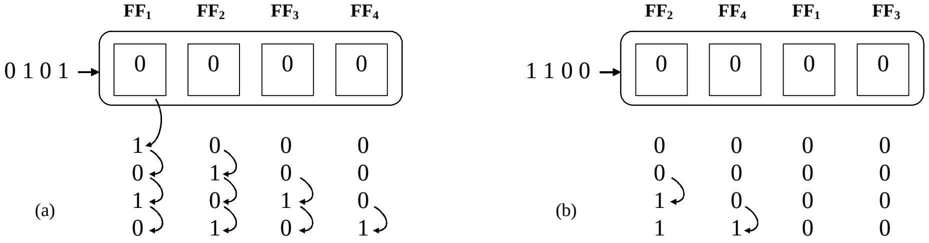
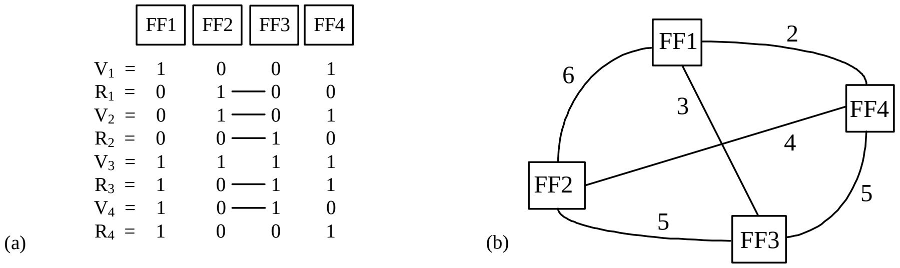


Figure 7.6: Impact of scan cell reordering on switching activity

A first scan-cell ordering technique was proposed in [Dabholkar 1998] which uses two heuristics to find the best ordering of scan cells. The first heuristic performs a random search where the scan cells are randomly permuted a predefined number of times and the entire deterministic test sequence is simulated to measure the switching activity. The second heuristic uses a **simulated annealing** algorithm that explores the whole space of solutions to search for a global optimum of the cost function. Both heuristics have met with limited success.

Another solution given in [Bonhomme 2002] starts from a set of scan cells and the deterministic test sequence generated to test the corresponding scan-based circuit. The method first constructs a complete undirected graph in which each vertex represents a scan cell and each edge represents a possible connection between two scan cells in the scan chain (see Figure 7.7b). The weight on each edge of the graph represents the total number of bit differences between two scan cells for the corresponding test sequence. In Figure 7.7a, V_i is a test vector, R_i is the corresponding output response, and there are 5 bit differences between scan cells FF2 and FF3 in this example. This weight reflects the number of transitions that may be generated in the corresponding portion of the scan chain by connecting these two scan cells together.



	FF1	FF2	FF3	FF4
V_1 =	1	0	0	1
R_1 =	0	1	0	0
V_2 =	0	1	0	1
R_2 =	0	0	1	0
V_3 =	1	1	1	1
R_3 =	1	0	1	1
V_4 =	1	0	1	0
R_4 =	1	0	0	1

Figure 7.7: An example test sequence and the corresponding weighted graph

From this weighted graph, the problem then amounts to finding a Hamiltonian path of minimum cost in the graph. The cost of a path is obtained by summing the weights on edges belonging to this path. This problem is equivalent to the well known traveling salesman problem, which is well known to be NP-hard and for which different polynomial-time approximation algorithms can be used. The solution implemented in [Bonhomme 2002] uses a greedy algorithm to find the scan cell ordering that minimizes the occurrence of transitions in the scan chain during scan-in and scan-out operations. The heuristic procedure can be exploited by any layout synthesis program during scan-cell placement and routing.

Scan-cell ordering has many advantages: (1) it does not require additional hardware, (2) the fault coverage and test time are left unchanged, and (3) the impact on the design flow is very low, and (4) significant reduction in test power can be obtained. The only drawback is that power-driven stitching of the scan cells may result in longer interconnect between the scan cells and potential routing congestion problems during scan routing (see Figure 7.8a).

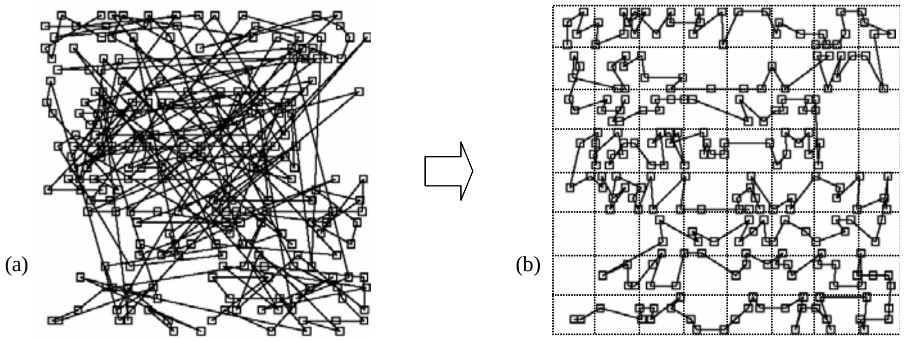


Figure 7.8: An example of power-driven scan chain routing

In order to provide better scan routing between scan cells after scan reordering for low power, the authors in [Bonhomme 2003] proposed to partition the circuit in clusters (by using geographical criteria) and then reorder the scan cells within each cluster so as to reduce the switching activity. The clusters are then stitched together using the nearest neighbor criteria. This technique offers a good tradeoff between test power reduction and scan chain length (see Figure 7.8b), and is applicable to circuits with multiple scan chains and clock domains.

7.4.6 Scan Architecture Modification

In this section, techniques that involve modifying the scan architecture by inserting new elements are presented.

A first solution involves partitioning the scan chain(s) into *N* segments and having only one segment at a time active when loading and unloading test data [Whetsel 2000]. An on-chip test module that contains a counter activates one segment at a time when it receives the scan enable signal from the ATE. When one segment has been completely loaded/unloaded, then the next segment is activated. This technique reduces the average power dissipated by a factor of *N*, without any change in the test time, the test sequence, or the fault coverage. Nevertheless, the power dissipation in the clock tree feeding the circuit, which represents a significant part of the total power dissipated during test, is not decreased. To address this, an alternative is to have separate clock tress for each scan segment so that the activation of the scan segments can be controlled by gating the clock trees rather than the scan enable signals [Saxena 2001]. In this way, the average test power is reduced in both the circuit and the clock tree without changing the overall principle of the method (see Figure 7.9).

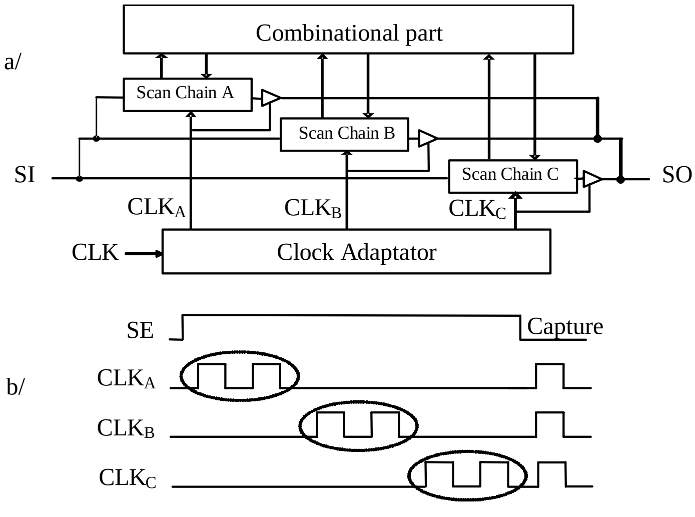


Figure 7.9: Scan chain segmentation

The same approach can be used to reduce peak power consumption during both shift operation and capture operation [Rosinger 2004]. A dedicated scan architecture with mutually exclusive scan segment activation is used for this purpose and a very high reduction in peak power consumption can be obtained.

Two other possible techniques based on scan architecture modification can be used [Sinanoglu 2002] [Lee 2000]. The first one [Sinanoglu 2002] consists of inserting logic elements (XOR gates) between the scan cells so as to minimize the occurrence of transitions in the scan chain (and hence in the CUT) during shift operations. Adding logic elements in the scan chains transforms the logic values that need to be shifted in. By doing this intelligently, it is possible to transform the scan vectors so that they contain fewer transitions. The second technique [Lee 2000] is applicable to circuits with multiple scan chains and consists of inserting different size buffers at each scan chain input to create a slight temporal skew between the scan chains during the shift operation. When a scan chain shifts, it creates transitions in the CUT which results in a current spike. By creating temporal skew between the shifting of the scan chains, the current spikes for each scan chain are spread out over time so that they do not occur simultaneously thereby allowing significant reductions in peak power during test. The technique requires some changes in the scan structure as well as in the scan controller.

Another interesting and original approach [Huang 2001] is based on a novel scan architecture, called a token scan architecture, that uses the concept of a token ring – a well-known structure in the field of communication networks – to reduce the shift power in both the scan chain and the combinational logic. Basically, this approach starts from a multiphase technique which is applied to scan-based circuits using the architecture shown in Figure 7.10.a. The scan-in wire *SI* is broadcasted to all scan cells but only one scan cell is activated at a time. For a scan chain with *N* scan cells, an *N*-phase non-overlapping clocking scheme is applied with one clock for each scan cell as shown in Figure 7.10.a. Since only one scan cell is activated at a time, a very high reduction of data transitions can be achieved during scan shifting.

However, the multiphase technique may have two problems due to the discrete multiphase generator. First, since the scan cells are usually distributed over the chip, the *N* multiphase clock routes will require large area overhead. Second, an inter-phase skew may occur due to the different lengths of the *N* clock routes, which may make multiphase clocks overlapped and cause a data error from *SI* or a bus contention at *SO*. To overcome these problems, a token ring-like structure [Huang 2001] can be used to embed the multiphase clock generator into the scan cell. As shown in Figure 7.10.b, a token bit is rotated in the scan chain and only the scan cell receiving the token can be activated. The *N* phase wires are thus reduced to a single phase clock wire. The inter-phase skew due to the different delays of the *N* phase wires can also be eliminated. This solution requires the use of a new type of scan cells, called token scan cells, to compose each scan chain (see Figure 7.11). A token scan cell consists of a data FF D_1, a token FF D_2, two multiplexers M_1 and M_2, and a switch S. D_1 and M_1 behave as a basic conventional scan cell while D_2 and M_2 serve as a phase generator.

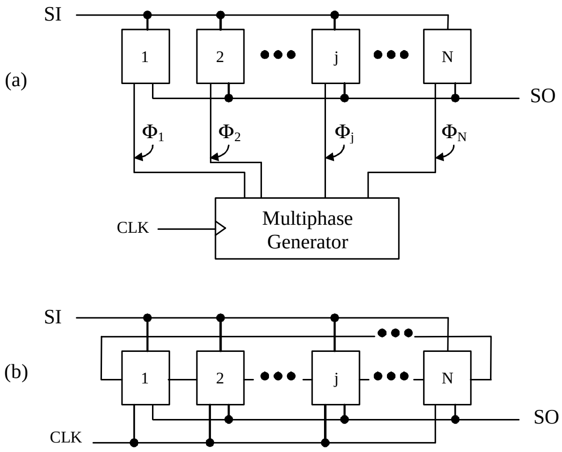


Figure 7.10: The token scan architecture

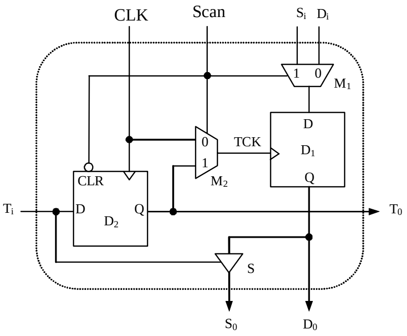


Figure 7.11: The token scan chain

In order to additionally reduce test power in the clock and scan-in data trees, the authors also propose a novel clock gating technique that takes the advantage of the regularity and periodicity of the token scan chain. By combining all the above concepts, the token scan architecture can efficiently reduce shift power as well as clock power - power consumed in the clock tree during scan testing. The main drawback of this solution is the significant area overhead.

7.4.7 Scan Clock Splitting

In order to reduce the power consumption during scan testing, it is also possible to modify the scan clock, i.e., the clock that drives all the scan cells of the chain (s). A first technique based on scan clock splitting [Bonhomme 2001] involves reducing the operating frequency of the scan cells during scan shifting without modifying the total test time. For this purpose, a clock whose speed is half of the normal (functional) clock is used to activate one half of the scan cells during one clock cycle of the scan operation (see Figure 7.12). During the next clock cycle, the second half of the scan cells in the scan chain(s) is activated by another clock whose speed is also half of the normal speed. The two clocks are synchronous with the system clock and have the same period during shift operation except that they are shifted in time. During capture operation, the two clocks operate as the system clock. The use of such a modified clock scheme lowers the transition density in the CUT, the scan chains and the clock tree feeding the scan chains during shift operation. Consequently, the switching activity in a time interval (i.e., the average power) as well as the peak power consumption is minimized. Moreover, the total energy consumption is also reduced as the test length with the proposed clock scheme is exactly the same than the test length with a conventional scan design to reach the same stuck-at fault coverage.

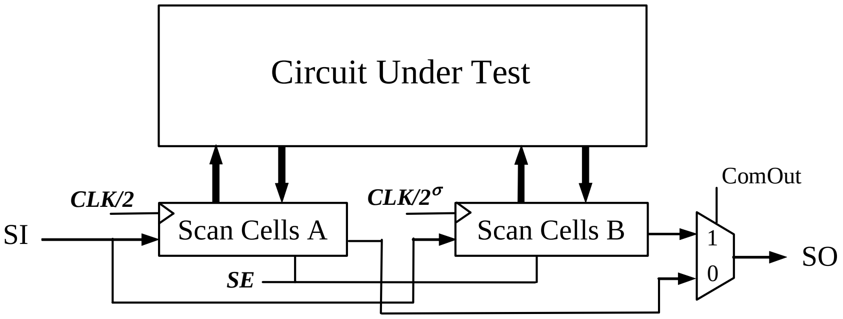


Figure 7.12: Scan clock splitting

Another technique [Sankaralingam 2003] uses a staggered clock scheme to reduce peak power dissipation during test. The principle of this approach is very similar to the one used in [Lee 2000]. The scan chains of the CUT are grouped together to form *N* groups of scan chains. Then, each clock cycle of the load/unload phase is divided in *N* periods, where each period corresponds to the activation of a given group of scan chains. By staggering the activation of each group in this manner, the number of scan cells that are simultaneously switching is reduced thereby greatly lowering the peak power consumption. Note, however, that the total number of transitions generated during test (i.e., the energy) is unchanged.

7.5 Low-Power Built-In Self-Test

Logic *built-in self-test* (BIST) is a DFT technique in which a portion of the ***circuit under test*** (CUT) is used to test itself. Because it can provide self-test ability, logic BIST is crucial in many applications, in particular, for safety-critical and mission-critical applications. One major objective of logic BIST is to obtain high fault coverage; however, a major issue is that power consumption during BIST can exceed the power rating of the chip or package. Increased average power can cause heating of the chip and increased peak power can produce noise-related failures [Bushnell 2000]. In this section, we discuss a number of low-power BIST architectures and methodologies to reduce power consumption.

7.5.1 Basics of Logic BIST

Figure 7.13 illustrates a typical logic BIST system. A **logic BIST controller** is required to control the BIST operation. The ***test pattern generator*** (TPG) automatically generates test patterns that are applied to the inputs of the ***circuit under test*** (CUT) and an ***output response analyzer*** (ORA) is used for compacting the circuit's output responses. In practice, in-circuit TPGs constructed from ***linear feedback shift registers*** (LFSRs) are commonly used for exhaustive, pseudo-exhaustive, or pseudo-random testing [Wang 2006]. This is mostly due to the fact that these LFSRs incur little area overhead and can be used as both TPG and ORA.

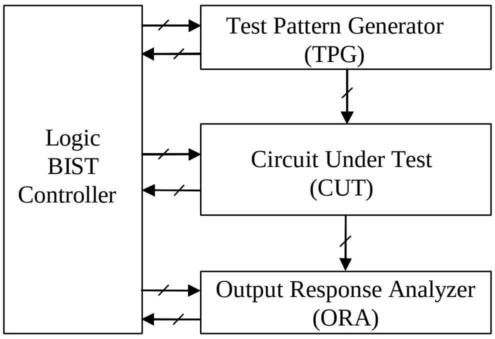


Figure 7.13: A typical logic BIST system

There are two basic BIST schemes for testing the circuit under test: (1) **test-per-clock BIST** for architectures using a register configurations, and (2) **test-per-scan BIST** for designs that incorporate scan chains [Wang 2006]. In test-per-clock BIST, test vectors are applied every clock cycle from the TPG, and test responses are captured in the ORA and compared to a reference value. This scheme has the advantage of running much faster in applying tests and yields higher fault coverage than test-per-scan BIST. The drawbacks of this scheme are high area overhead and incompatibility with scan design. In test-per-scan BIST, also called **scan-based BIST**, pseudo-random patterns are first shifted into the scan chains during shift operation; the test responses to these patterns are then captured in scan cells during the capture operation. The captured test responses are shifted out to the ORA for response compaction while a new test is being shifted in. Clearly, the test-per-scan BIST system will run much slower than the test-per-clock BIST system; however, it takes advantage of the existing scan design thereby requiring much simpler BIST circuitry and is thus the industry preferred solution today.

7.5.2 LFSR Tuning

The aim of LFSR tuning is to find a way of decreasing the energy consumed during BIST by appropriately selecting the parameters of the LFSR, i.e., the seed and characteristic polynomial. A preliminary step in this approach is to analyze the impact of these parameters on the switching density generated in the CUT. In [Girard 1999a], a number of experiments on benchmark circuits were conducted where for each circuit, several characteristic polynomials were used for the LFSR, and for each of these polynomials, several seeds were tried. Polynomials were taken from the list of primitive polynomials of an *n*-stage LFSR (*n* being the number of primary inputs of the CUT), and seeds were randomly chosen for each selected polynomial. In each experiment, the length of the test sequence required to reach the target fault coverage was determined through fault simulation. In Figure 7.14, the experimental results for an 8 by 8 multiplier targeting 99% stuck-at fault coverage are shown. Each number on the X axis corresponds to a particular primitive polynomial of the LFSR, and each dot corresponds to the internal WSA resulting from a randomly selected seed for the particular polynomial. Note that the *internal* WSA refers to the **weighted switching activity** of the internal nodes of the CUT.

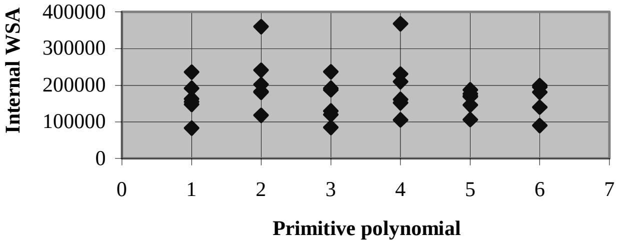


Figure 7.14: Impact of LFSR polynomial selection on energy

As can be seen in the figure, the WSA obtained for a given primitive polynomial of the LFSR strongly depends on the seed selected. Indeed, the deviation between best seeds and worst seeds is very significant in terms of WSA. On the other hand, sensitivity of the WSA to a given primitive polynomial is much lower; the value of the minimum WSA is almost the same regardless of which primitive polynomial is used. Therefore, selecting a primitive polynomial to minimize energy dissipation during BIST is not as crucial as selecting a good seed for the LFSR. For a given polynomial and target fault coverage, selecting the best seed of an LFSR for low-power BIST can then easily be done by using a method based on a simulated annealing algorithm [Girard 1999a].

7.5.3 Low-Power Test Pattern Generators

Several approaches have been proposed for designing on-chip test generators that can generate effective test patterns while reducing the transition density in the CUT. A first approach, called ***dual speed LFSRs*** (DS-LFSRs) [Wang 1997], is based on the use of two LFSRs operating at different clock frequencies (see Figure 7.15). Average power during test is reduced by connecting the CUT inputs with the highest transition densities to the low speed LFSR while CUT inputs with the lowest activity are connected to the normal speed LFSR. Note that this technique is applicable in a test-per-clock BIST environment. A second approach is also based on a modified LFSR [Girard 2001]. The original LFSR is replaced by two LFSRs that operate out-of-phase at half the clock rate of the original (functional) speed. Compared to the previous approach, the power dissipation is reduced not only in the CUT but also in the clock tree feeding the circuit. Fault coverage and test time are left unchanged. A third solution [Zhang 1999] consists of inserting logic between the LFSR and the CUT to allow the generation of weighted random test patterns that reduce the switching activity in the circuit while maintaining a high fault coverage. This solution uses a genetic algorithm based search to determine optimal weight sets at primary inputs to minimize energy dissipation. One last approach that can be used [Corno 2000] is based on selecting a cellular automata that generates a test sequence with a low transition density and has a good tradeoff between fault coverage and test time.

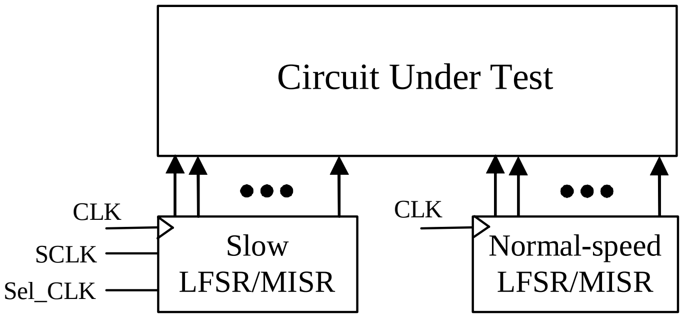


Figure 7.15: Low-power BIST with DS-LFSRs

For scan-based BIST (test-per-scan BIST), an interesting approach for low-power testing, called ***low transition random test pattern generator*** (LT-RTPG) [Wang 1999], was proposed. It involves inserting an AND gate and a *toggle flip-flop* (TFF) between the LFSR and the input of the scan chain, so as to increase the correlation of neighboring bits in the scan vectors (see Figure 7.16). Since the TFF holds its previous values until it receives a 1 on its input, the same value (0 or 1) is repeatedly scanned into the scan chain until the value at the output of the AND gate becomes 1. Hence, neighboring scan cells are assigned identical values in most test vectors if a large *k* is used, i.e., if the AND gate has many inputs (the probability that the TFF toggles at any time *t* is given by $1/2^k$). In this manner, the number of transitions generated in the CUT can be significantly reduced. Although the pseudo-random test sequence is modified by this logic, it still provides a good tradeoff between fault coverage and test time.

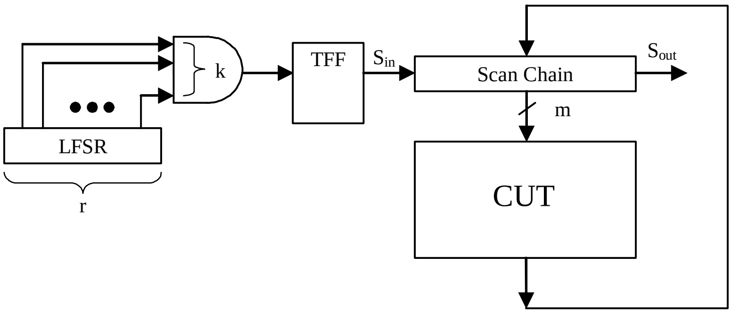


Figure 7.16: The LT-RTPG structure

Interesting low-power test pattern generators that are applicable for data path architectures based on multipliers and accumulators are described in [Gizopoulos 2000]. Two hardware solutions are proposed depending on whether the concern is energy reduction or power reduction. They are both based on the use of Gray counters which can generate successive test vectors with a Hamming distance of one. Significant energy and power reductions can be obtained.

7.5.4 Vector Filtering BIST

BIST techniques based on vector filtering to reduce power consumption have also been proposed in the literature [Girard 1999b] [Manich 2000] [Corno 1999] [Gerstendörfer 1999]. These techniques are based on the observation that as self-test progresses, the detection capability of the pseudo-random test vectors generated by an LFSR decreases very quickly. Therefore, many of the pseudo-random test vectors do not detect new faults despite consuming a significant amount of energy. For example, only 159 patterns among the 2524 required to reach 99.9% fault coverage actually detect faults in the benchmark circuit c5315 [Brglez 1985]. In addition, the length of the sub-sequence of consecutive non-detecting test vectors is often long. For example, the longest sub-sequence of consecutive non-detecting vectors in the pseudo-random test sequence generated for the benchmark circuit s1488 [Brglez 1985] to reach 100% fault coverage contains 509 vectors, while the complete test sequence is of length 2931.

Consequently, the main goal of these techniques is to filter test vectors that do not detect additional faults, thus preventing the CUT from being excited by these undesired vectors. For this purpose, a decoder can be used during BIST pattern generation to store the first and last vectors of each sub-sequence of consecutive non-detecting vectors to be filtered [Girard 1999b]. The output of this decoder provides the logic value 1 after detection of each of these vectors. Then, the vector filtering structure has to allow or prevent application of these test vectors to the circuit inputs. A toggle D flip-flop is used to control the transmission of stimuli from the LFSR to the CUT. The transmission is activated or inhibited by means of a transmission gate network (see Figure 7.17a).

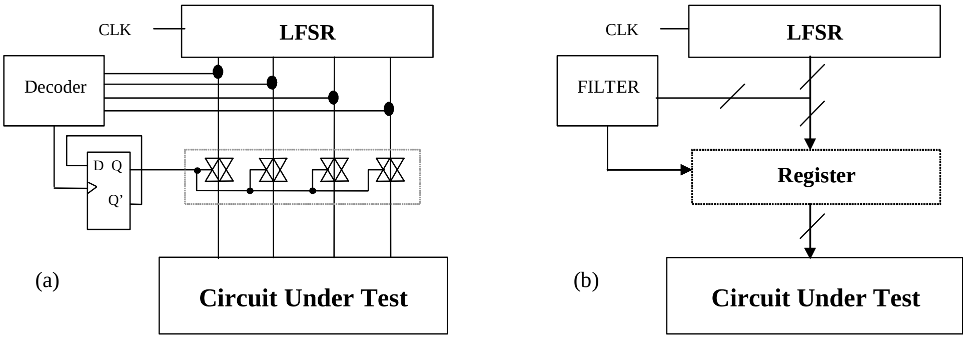


Figure 7.17: Two vector filtering BIST structures

Rather than just filtering the sub-sequences of non-detecting vectors, it is also possible to filter all vectors that do not detect any faults [Manich 2000]. Using this approach, a register made of latches is used to control the transmission of test vectors from the LFSR to the CUT. A filter module is used to provide the needed control signals to the register (see Figure 7.17b). A similar solution was also proposed in [Corno 1999].

The authors in [Gerstendörfer 1999] exploited the same idea but apply it to scan-based BIST. A gating signal is derived from the decoder/filter and is used to enable or disable the shift clock to the scan chains.

The main advantage of all these techniques is that they allow significant reduction of energy and average power consumption during testing. The drawbacks are the negative impact on circuit performance and the area overhead which may be high in some cases.

7.5.5 Circuit Partitioning

Another approach involves partitioning the original circuit into two structural sub-circuits so two different BIST sessions can be used to successively test each sub-circuit. In order to minimize the area overhead of the resulting BIST scheme, however, the number of connections between the sub-circuits of the partition, called the *cut size*, has to be minimal. The basic scheme of partitioning a circuit into two sub-circuits is shown in Figure 7.18. In Figure 7.18a, a logic circuit is partitioned into two sub-circuits C_1 and C_2. Many such partitions exist for large VLSI circuits. Figure 7.18b depicts how multiplexers are inserted between the two sub-circuits. By controlling the multiplexers, all inputs and outputs of each sub-circuit can be accessed using primary inputs and primary outputs. For example, to test sub-circuit C_1, the multiplexers can be controlled as shown in

Figure 7.18c. The demultiplexers (DMUX) on the sets *B* and *C* of input signals are added to avoid switching activity in C_2 during the test of C_1.

A circuit partitioning technique for low power testing was proposed in [Girard 1999c]. This technique tries to find an optimal partitioning solution, which is a *NP*-complete problem, by using a simple *graph partitioning* algorithm. An improved version [Girard 2000] uses a circuit partitioning tool based on a *multilevel hypergraph partitioning* algorithm [Karypis 1998]. Traditional partitioning algorithms compute a partition of a graph by directly operating on the original graph. This approach is often too slow and can lead to poor quality partitions in terms of cut size which is representative of the area overhead of the BIST scheme. The multilevel partitioning algorithm follows a completely different approach. The algorithm successively decreases the size of the graph (or the **hypergraph**) by collapsing vertices and edges, partitions the smallest graph, and then uncoarsens it to construct a partition for the original graph. At each level of the uncoarsening phase, an iterative refinement algorithm is used to improve the partition.

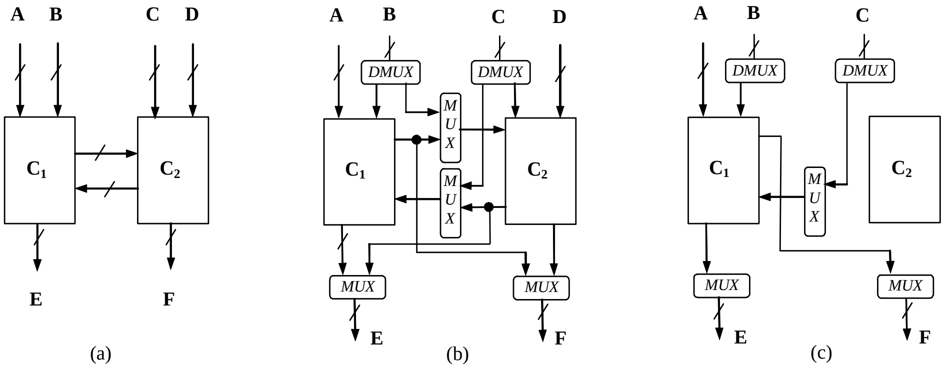


Figure 7.18: Circuit partitioning for low-power testing

By partitioning the circuit into two sub-circuits and testing the sub-circuits in successive test sessions, average and peak power consumption are minimized. In addition, this approach reduces the total energy consumed during BIST operation because the test length required for the two sub-circuits is usually shorter than that of the original circuit. This is due to the fact that circuit partitioning increases the controllability and observability of the internal nodes in the CUT. The area overhead with this approach is low. Drawbacks are a slight penalty on circuit performance and a non-negligible impact on routing. The proposed strategy can be applied to scan-based BIST or parallel BIST by adapting the test pattern generation structure.

7.5.6 Power-Aware Test Scheduling

A test scheduling technique for low power consumption [Zorian 1993] considers a set of blocks (memories, logic, analog, test resources, etc.) in an SOC and a specified limit of power dissipation for the SOC during test. The objective is to find the best combination of blocks to be tested in parallel so the overall test time is minimal and the power limit is satisfied. This technique also takes into account the fact that, in order to minimize the area overhead associated to BIST, some of the test resources (test pattern generators and output response analyzers) must be shared among the various blocks.

A similar technique [Chou 1994] addresses the *NP*-complete test scheduling problem by using a compatibility graph and heuristic-driven algorithms. The power constraint is established with respect to the peak power consumption of each block. Two different problems are considered depending on the test length of each block: 1) scheduling equal-length tests with power constraints and 2) scheduling unequal-length tests with power constraints. Optimal solutions are sought for both problems. The algorithms consist of four basic steps. First, a test compatibility graph is constructed from a resource graph in which a resource represents either a combinational block or a register block. Second, the test compatibility graph is used to identify a complete set of **time compatible tests** (tests that can be executed concurrently) with power dissipation information associated with each test. Third, from the set of time compatible tests, lists of **power compatible tests** are extracted. Finally, a minimum cover approach is used to find an optimum scheduling of these power compatible tests.

Based on the above two basic test scheduling techniques, several solutions have been further proposed for testing SOC designs [Muresan 2000] [Iyengar 2001] [Larsson 2002] [Pouget 2003]. For given power constraints and parameters related to the test organization (fixed, variable, or undefined test sessions with or without precedence constraints) or to the test structure (test bus width, test resources sharing), these solutions allow to optimize overall SOC test time.

Another test scheduling technique [Ubar 2005] has a slightly different objective, as the main focus is on total test energy minimization for SOC testing. This technique assumes a hybrid BIST test architecture, where the test set is composed of core-level locally generated pseudo-random test patterns and additional deterministic test patterns that are generated off-line and stored in the system (see Figure 7.19). The exact composition of these patterns defines not only the test length and test memory requirements, but also the energy consumption. In general, since a deterministic test pattern is more effective in detecting faults than a pseudo-random pattern, using more deterministic test patterns for a core will lead to a short test sequence with, consequently, less energy. However, the total number of deterministic test patterns is constrained by the test memory requirements, and at the same time, the deterministic test patterns of different cores of a SOC have different energy and fault detection characteristics. A careful tradeoff between the deterministic pattern lengths of the core must therefore be made in order to produce a globally optimal solution. Two heuristics [Ubar 2005] can be proposed to try to minimize the total switching energy without exceeding the assumed test memory constraint. The solutions are obtained by modifying the ratio of pseudo-random and deterministic test patterns for every individual core such that the total energy dissipation is minimized.

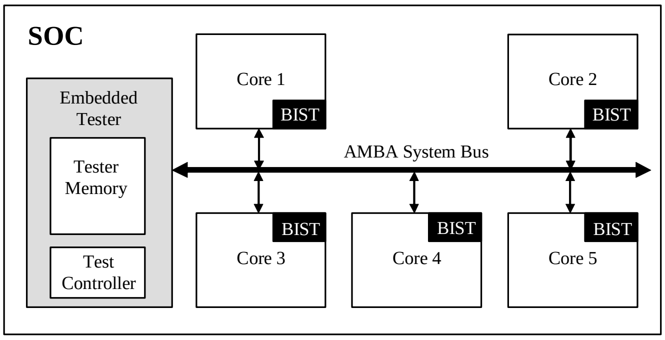


Figure 7.19: AMBA bus-based hybrid BIST architecture

Another category of test scheduling, called thermal-aware test scheduling, has been proposed to address the problem of chip overheating during test of complex core-based systems. Here, the basic idea is to consider that the spatial distribution of power across the chip is non-uniform, so that imposing a chip-level maximum power constraint during test scheduling (as for system-level BIST solutions described above) doesn’t necessarily avoid local overheating and hence destructive hot spots. A few solutions have been proposed in this area [Rosinger 2005] [Liu 2005] [He 2006], mainly based on incorporating thermal constraints during test scheduling so as to spread heat more evenly over the chip and reduce hot spots during test. The proposed approaches facilitate rapid generation of thermal-safe test schedules without requiring time-consuming thermal simulations.

7.6 Low-Power Test Data Compression

Test data volume is now recognized as a major contributor to the cost of manufacturing testing of ICs. High test data volume leads to a high testing time and may exceed the limited memory depth of Automatic Test Equipment (ATE). Test application time for scan testing can be reduced by using a large number of scan chains. However, the number of ATE channels that can directly drive scan chains is limited due to pin count constraints.

Test data compression is an efficient solution to the problem of increasing test data volume. Test data compression involves encoding a test set so as to reduce its size. By using this reduced set of test data, the ATE limitations, i.e., tester storage memory and bandwidth gap between the ATE and the CUT, may be overcome. On the other hand, using compressed test data involves having a small on-chip decoder which decompresses the data as it is fed into the scan chains during test application.

Despite its ability in reducing test data volume and test application time, test data compression does not solve the problem of excessive test power during scan testing. A case study of a Motorola ColdFire® microprocessor core [Pouya 2000] has been used to illustrate the commercial means of reducing test data volume and how they affect test power. To address this issue, several techniques have been proposed to simultaneously reduce test data volume and test power during scan testing of digital ICs. As in [Wang 2006], these low-power test data compression techniques can be classified into three categories: **coding-based** schemes, **linear-decompression-based** schemes, and **broadcast-scan-based** schemes.

7.6.1 Coding-Based Schemes

Code-based schemes use data compression codes to encode the test cubes of a test set. An interesting encoding algorithm that can be used to concurrently reduce scan power dissipation and test data volume during SOC testing is proposed in [Chandra 2001]. In this approach, test cubes generated by ATPG are encoded using **Golomb codes** which are an evolved form of **run-length codes**. All don’t care bits of the test cubes are mapped to 0 and Golomb coding is used to encode runs of 0’s. More details about Golomb codes can be found in [Wang 2006]. Golomb coding efficiently compresses test data, and the mapping of don’t cares to all 0’s reduces the number of transitions during scan-in, thus significantly reducing power dissipation (up to 75%). One drawback of Golomb coding is that it is very inefficient for runs of 1’s. In fact, the test storage can even increase for test cubes that have many runs of 1’s. Moreover, implementing this test compression scheme requires a synchronization signal between the ATE and the CUT as the size of the compressed data (*codeword*) is of variable length.

Another method based on an **alternating run-length coding** [Chandra 2002] improves the encoding efficiency of Golomb coding. While a Golomb code only encodes runs of 0’s, an alternating run-length code can encode both runs of 0’s and runs of 1’s. In this case, the drawback is that the coding becomes inefficient when a pattern with short runs of 0’s or 1’s has to be encoded.

7.6.2 Linear-Decompression-Based Schemes

Another class of low-power test stimulus compression schemes is based on using **linear decompressors** to expand the data coming from the tester to fill the scan chains during test application. Linear decompressors consist only of XOR gates and flip-flops, and are described in details in [Wang 2006].

An example of a low-power linear-decompression-based scheme using LFSR reseeding is proposed in [Lee 2004]. The basic idea in LFSR reseeding is to generate deterministic test cubes by expanding seeds. A seed is an initial state of the LFSR that is expanded by running the LFSR in an autonomous mode. Given a deterministic test cube, a corresponding seed can be computed by solving a set of linear equations – one for each specified bit – based on the feedback polynomial of the LFSR. Since typically 1-5% of the bits in a test vector are specified, most bits in a test cube do not need to be considered when a seed is computed because they are don't care bits. Therefore, the size of a seed is much smaller than the size of a vector. Consequently, reseeding can significantly reduce test data volume and bandwidth. However, it is not as good for power consumption because the don't care bits in each test cube get filled with random values thereby resulting in excessive switching activity during scan shifting.

The key idea of the encoding scheme proposed in [Lee 2004] is to take advantage of the fact that the number of transitions in a test cube is always less than its number of specified bits. A transition in a test cube is defined as a specified 0 (1) followed by a specified 1 (0) with possible *X*'s between them, *e.g.*, X10XXX or XX0X1X. Thus, rather than using LFSR reseeding to directly encode the specified bits as in conventional LFSR reseeding, the proposed encoding scheme divides each test cube into blocks and only uses LFSR reseeding to produce the blocks that contain transitions. For the blocks that do not contain transitions, the logic value fed into the scan chain is simply held constant. This approach reduces the number of transitions in the scan chain and hence reduces test power. Despite the area overhead due to the use of *hold flag* shift registers, this scheme is an efficient solution to trade-off between test data compression and test power reduction.

7.6.3 Broadcast-Scan-Based Schemes

The third class of low-power test data compression schemes is based on broadcasting the same value to multiple scan chains.

An example of a low-power broadcast-scan-based scheme is the **segmented addressable scan** architecture presented in [Al-Yamani 2005]. This architecture involves modifying the **Illinois scan architecture** [Hamzaoglu 1999] in which a given scan chain is split into multiple scan segments thus allowing the same data to be loaded simultaneously into all segments when compatibility exists. The segmented addressable scan architecture enhances the Illinois scan architecture by avoiding the limitation of having to have all segments compatible to benefit from the segmentation. In other words, any combination of compatible segments for a given test pattern can be used to load the same data to these segments and hence increase the compression rate. The compatible segments are loaded in parallel using a multiple-hot decoder (see Figure 7.20). Test power is reduced as segments which are incompatible within a given round, *i.e.*, during the time needed to upload a given test pattern, are not clocked. One drawback of this solution is that the multiple-hot decoder is designed with respect to a given test set. This means that the test set has to be known early during the design phase of the circuit, and that changing the test set during verification test or production test involves changing the design of the circuit.

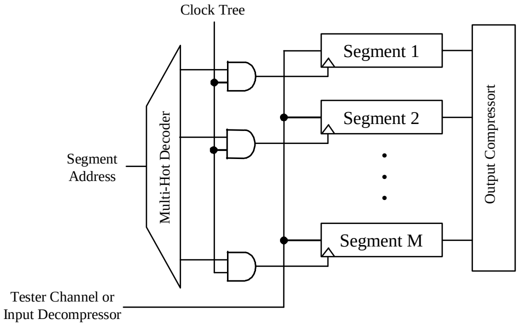


Figure 7.20: The segmented addressable scan architecture

Another example is the ***progressive random access scan*** (PRAS) architecture proposed in [Baik 2005] that allows individual accessibility to each scan cell. In this architecture, scan cells are configured as an SRAM-like grid structure using specific PRAS scan cells and some additional peripheral and test control logic (see Figure 7.21). Providing such accessibility to every scan cell eliminates unnecessary switching activity during scan, while reducing the test application time and test data volume by updating only a small fraction of scan-cells throughout the test application. Power consumption during test is drastically reduced - up to 99%. The main drawback of the PRAS architecture is the significant hardware overhead.

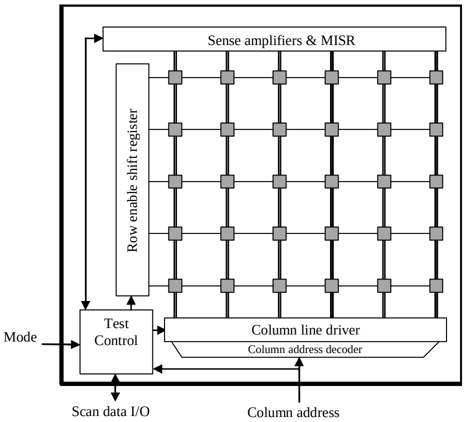


Figure 7.21: The progressive random access scan (PRAS) architecture

7.7 Low-Power RAM Testing

While numerous techniques for constraining power dissipation during test exist, there appears to be only a few solutions that are dedicated to memories. The main motivation for reducing test power in memories can be explained as follows. System memories, or embedded device memories, are divided into banks for increasing access speed and optimizing system cost [Cheung 1996]. During normal system operation, only one memory bank is accessed at any given time. In contrast, concurrent self-test of all memory modules is highly desirable to reduce test time and simplify BIST control circuitry. However, by concurrently testing several banks of memories, the power dissipation can by far exceed that during normal system operation. For this reason, reducing test power in memories becomes mandatory when concurrent testing is used. Note that this statement also applies when testing memories embedded in an SOC.

A first methodology for low-power test of *Random Access Memories* (RAMs) [Cheung 1996] is based on modifying several common memory tests (Zero-One, Checker Board, March B, Walking-0-1 and SNP 2-Group) in a way that reduces power dissipation during test. The modified tests are based on the following principle: reorder the original tests to minimize the switching activity on each address line while retaining the fault coverage. The number of signal transitions on each address line depends on the address counting method (i.e., the order in which addresses are enumerated during a read or a write loop of a memory test algorithm) and the address bit position. For example, the LSB (respectively MSB) address line has the largest (respectively smallest) number of transitions when binary address counting is used. Table 7.1 describes the original and low-power test algorithms for two memory tests. The symbol ‘$\updownarrow$’ is used to describe a sequential access to all memory cells in any addressing order (increasing or decreasing). Binary address counting is typically used for such addressing. The low-power tests are described using the symbol $\updownarrow_s$ which represents *Single Bit Change* (SBC) counting. For example, {00, 01, 11, 10} is the counting sequence of a two bit SBC code. Finally, W0 (W1) represents writing a 0 (1) to an address location and R0 (R1) represents reading a 0 (1) from an address location.

Table 7.1: Original and low-power memory test algorithms

	Original Test	**Low-power Test**
Zero-One	$\updownarrow$ (W0); $\updownarrow$ (R0); $\updownarrow$ (W1); $\updownarrow$ (R1);	$\updownarrow_s$ (W0, R0, W1, R1);
Checker Board	$\updownarrow$ (W($1_{odd}/0_{even}$)); $\updownarrow$ (R($1_{odd}/0_{even}$)); $\updownarrow$ (W($0_{odd}/1_{even}$)); $\updownarrow$ (R($0_{odd}/1_{even}$));	$\updownarrow_s$ (W($1_{odd}/0_{even}$), R($1_{odd}/0_{even}$), W($0_{odd}/1_{even}$), R($0_{odd}/1_{even}$));

Each proposed test has the same fault coverage and time complexity as the original version, but reduces power dissipation by a factor of two to sixteen thanks to a modified addressing sequence. A special design of the BIST circuitry [Cheung 1996] is required to implement the proposed low-power tests.

Another methodology [Dilillo 2006] to minimize test power in SRAM memories is to exploit the predictability of the addressing sequence. It is known that the pre-charge circuits are the principal contributor to power dissipation in SRAM. It has indeed been shown that it may represent up to 70% of the overall power dissipation of an SRAM memory [Liu 1994]. These circuits have the role of pre-charging and equalizing the long and high capacitive bit lines. This action is essential to ensure correct memory operation. To reduce the pre-charge activity during test, one can use the fact that in functional mode the cells are selected in random sequence, and therefore all pre-charge circuits need to be always active, while during the test mode the access sequence is known, and hence only the columns that are to be selected need to be pre-charged [Dilillo 2006]. This low-power test mode can be implemented by using a modified pre-charge control circuitry, and by exploiting the first degree of freedom of March tests, which allows choosing a specific addressing sequence. The modified pre-charge control logic contains an additional element for each column (see Figure 7.22). This element consists of one multiplexer (two transmission gates and one inverter) and one NAND gate. Signal LP_{test} allows the selection between the functional mode of the memory and the low-power test mode in which the addressing sequence is fixed to "word line after word line" and the pre-charge activity is restricted to two columns for each clock cycle: the selected column and the following one. Signal Pr_j is the pre-charge signal originally used, while CS_j' is the complement of the column selection signal. The multiplexer operates the mode selection, while the NAND gate forces the functional mode for a given column when it is selected for a read/write operation during test. When LP_{test} is ON, the signal CS_j' of a column *j* drives the pre-charge of the next column *j+1*. Note that the pre-charge is active with the input signal at '0'.

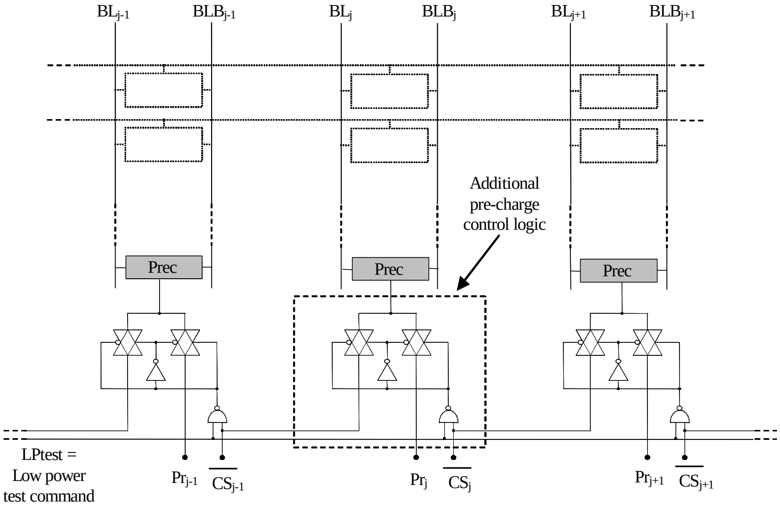


Figure 7.22: A pre-charge control logic for low-power testing

Experiments used to validate the proposed method have shown a significant test power reduction (~50%) with negligible impact on area overhead and memory performance.

7.8 Summary and Conclusions

Numerous studies from academia and industry, including the *International Technology Roadmap for Semiconductors* (ITRS) published by the Semiconductor Industry Association (SIA) [SIA 2005] [SIA 2006], have shown the need to reduce power consumption during test of digital and memory designs. This need is triggered by the fact that typically test power can be more than twice the power consumed in normal functional mode.

Because test throughput and manufacturing yield are often affected by test power, various test solutions have been proposed over the past decade. In this chapter, we discuss many low-power test solutions to address the above-mentioned problems. Both structural and algorithmic solutions are described along with their impacts on parameters such as fault coverage, test time, area overhead, circuit performance penalty, and design flow modification. These solutions cover a broad spectrum of testing environments, including scan testing, scan-based BIST, test-per-clock BIST, test compression, and memory testing.

While solutions presented in this chapter can be used to address most of the problems caused by excessive test power, not all problems have been solved. One concern is when multiple issues arise at the same time when developing low-power test solutions. For example, almost all digital circuits today have scan chains, and quite a few require test compression for test data volume reduction along with at-speed testing for screening timing defects. Thus far, few solutions have been proposed to address the problem of low-power scan testing when both test compression and at-speed scan testing are used. Another concern is related to the growing complexity and increasing use of core-based systems. In this case, we are now facing situations where several cores, such as scan cores, memory cores, and logic BIST cores each with embedded at-speed test features, have to be tested in parallel to avoid prohibitive test time. Power-aware or thermal-aware test scheduling is required for these SOC designs so that power and thermal constraints are satisfied while maintaining an optimized test throughput. This complicates the low-power test problems and requires the joint development of core-level and system-level low-power test solutions in the nanometer SOC design era.

Finally, concerns arise from how testing is to be done when new low-power design techniques, such as dynamic power management and multiple-voltage design techniques, are employed. The idea of dynamic power management is to "shut-down" parts of a design when they are idle. Thus far, testing of those designs has been done sequentially, i.e., dealing with power domains one at a time. However, this practice will soon become inadequate due to test time concern. Similarly, multiple-voltage domains have also been used in designs to reduce power consumption. Among others, the challenges we are facing now include how to build scan chains that span more than one voltage domain, cope with physical design constraints when dealing with level-shifters, and how to safely handle the test of such designs. A more future issue relates to asynchronous design, which is now seeing renewed interest as a way to reduce power. Although still far from being practical, asynchronous design will also require new and dedicated low-power test solutions.

7.9 Exercises

7.1 (**Test Power)** Provide at least three examples to show why scan test power can be significantly higher than functional power.

7.2 (**Test Power Reduction)** List three *ad hoc* solutions for reducing power consumption during test application, and show the advantages and disadvantages of each solution.

7.3 (**Terminology)** Explain the difference between dynamic power, short-circuit power, and leakage power.

7.4 (**Terminology**) Explain the difference between energy and power. Also explain what the following terms mean: average power, instantaneous power, and peak power.

7.5 (**Test Power Evaluation**) Show the equations for estimating average power, instantaneous power, and peak power. Explain all the parameters used in the equations.

7.6 (**Noise Phenomena)** Describe three types of circuit noise and their impact on testing.

7.7 (**Terminology**) Explain the following terms:

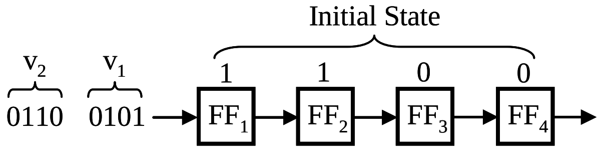
(1) shift, test, capture cycles in scan testing
(2) slow-speed scan testing, at-speed scan testing
(3) launch-on-shift (LOS), launch-on-capture (LOC)

7.8 (**Test-Induced Yield Loss)** Give at least three reasons why yield loss may occur due to scan testing.

7.9 (***X*-Filling)** Conduct *minimum transition filling* (MT-filling), 0-filling, and 1-filling for the following test cube and calculate the number of weighted transitions for each resulting fully-specified test vector:

Scan-Input-Pin: 1XX0XXXX0XXXXX101XX0

7.10 (**Scan Cell Ordering)** Consider the application of two test vectors to a scan chain composed of 4 scan cells as follows:



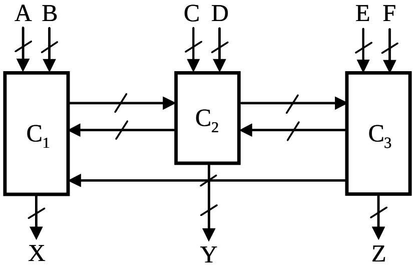
Find the best order of scan cells for reducing the average switching activity measured by the number of weighted transitions.

7.11 (**Token Scan Architecture)** Consider the token scan scheme shown in Figure 7.10. This scheme allows only one scan cell to be activated each time. Show a new token scan scheme that allows exactly two scan cells to be activated each time.

7.12 (**Scan Clock Splitting)** Consider the scan clock splitting scheme shown in Figure 7.12. This scheme divides scan cells into two groups that are not operated simultaneously. Show a new scan clock splitting scheme that divides scan cells into three groups that are not operated simultaneously.

7.13 (**Test Vector Filtering BIST**) Explain the basic idea of test vector filtering BIST, and then show two possible techniques to implement this idea.

7.14 (**Circuit Partitioning)** Study the example (Figure 7.18) of circuit partitioning for low-power testing in logic BIST. Then show how to partition the following circuit for low-power testing.



7.15 (**Low-power RAM Testing)** Compare the original test algorithms and low-power test algorithms shown in Table 7.1. Discuss the reasons why the low-power test algorithms can reduce test power.

7.16 (**A Design Practice)** Use the ATPG programs and user's manuals contained on the Companion Web site to generate test sets for a number of full-scan benchmark circuits, with and without using the low shift power option. Compare the resulting test sets in term of fault coverage, test data volume, and estimated shift power dissipation.

7.17 (**A Design Practice)** Use the ATPG programs and user's manuals contained on the Companion Web site to generate test sets for a number of full-scan benchmark circuits, with and without using the low capture power option. Compare the resulting test sets in term of fault coverage, test data volume, and estimated capture power dissipation.

7.18 (**A Design Practice)** Use the ATPG programs and user's manuals contained on the Companion Web site to generate test sets for a number of full-scan benchmark circuits, with and without using the low shift/capture power option. Compare the resulting test sets in term of fault coverage, test data volume, and estimated capture power dissipation.

Acknowledgments

The authors wish to thank Prof. Nicola Nicolici of McMaster University in Canada, Dr. Mokhtar Hirech of Synopsys Inc. in USA, and Dr. Christian Landrault of LIRMM/CNRS in France, for reviewing this chapter and providing helpful comments. They also wish to thank Dr. Zhi G. Jiang of SynTest Technologies, Inc. for providing exercise questions on Design Practices.

References

R7.0 Books

[Altet 2002] J. Altet and A. Rubio, *Thermal Testing of Integrated Circuits*, Springer Science, New York, NY,2002.

[Bushnell 2000] M. L. Bushnell and V. D. Agrawal, *Essentials of Electronic Testing for Digital, Memory & Mixed-Signal VLSI Circuits*, Springer Science, New York, NY, 2000.

[Crouch 1999] A. Crouch, *Design-For-Test for Digital IC's and Embedded Core Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[Krstic 1998] A. Krstic and K.-T. Cheng, *Delay Fault Testing for VLSI Circuits*, Springer Science, New York, NY, 1998.

[Nicolici 2003] N. Nicolici and B. Al-Hashimi, *Power-Constrained Testing of VLSI Circuits*, Springer Science, New York, NY, 2003.

[Rajski 1998a] J. Rajski and J. Tyszer, *Arithmetic Built-In Self-Test for Embedded Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[Wang 2006] L.-T. Wang, C.-W. Wu, and X. Wen, *VLSI Test Principles and Architectures: Design for Testability,* Morgan Kaufmann, San Francisco, CA, 2006.

[Weste 1993] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective,* Second Edition, Addison-Wesley, 1993.

R7.1 Introduction

[De Colle 2005] A. De Colle, S. Ramnath, M. Hirech and S. Chebiyam, "Power and Design for Test: A Design Automation Perspective," *ASP Journal of Low Power Electronics*, Vol. 1, No. 1, pp. 73-84, April 2005.

[Girard 2000] P. Girard, "Low Power Testing of VLSI Circuits: Problems and Solutions," *Proc. Int'l Symp. on Quality of Electronic Design,* pp. 173-179, March 2000.

[Monzel 1997] J. Monzel, S. Chakravarty, V. D. Agrawal, R. Aitken, J. Braden, J. Figueras, S. Kumar, H.-J. Wunderlich, and Y. Zorian, "Power Dissipation During Testing: Should We Worry About it?," *IEEE VLSI Test Symp.*, Panel Session, April 1997.

[Pouya 2000] B. Pouya and A. Crouch, "Optimization Trade-offs for Vector Volume and Test Power," *Proc. Int'l Test Conf.*, pp. 873-881, October 2000.

[Saxena 2003] J. Saxena, K. M. Butler, V. B. Jayaram, S. Kundu, N. V. Arvind, P. Sreeprakash, and M. Hachinger, "A Case Study of IR-Drop in Structured At-Speed Testing," *Proc. Int'l Test Conf.*, pp. 1098-1104, October 2003.

[SIA 2001] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors: 2001 Edition*, http://public.itrs.net, 2001.

[Wang 1997] S. Wang and S. K. Gupta, "DS-LFSR: A New BIST TPG for Low Heat Dissipation," *Proc. Int'l Test Conf.*, pp. 848-857, November 1997.

[Zorian 1993] Y. Zorian, "Testing the Monster Chip," *IEEE Spectrum*, Vol. 36, No. 7, pp. 54-60, July 1999.

R7.2 Energy and Power Modeling

[Athas 1994] W.C. Athas, L.J. Svensson, J.G. Koller, N. Tzartzanis and E. Ying-Chin Chou, "Low-Power Digital Systems Based on Adiabatic-Swticging Principles," *IEEE Trans. on Very Large Scale Integration Systems*, Vol. 2, No. 4, pp. 398-416, Dec. 1994.

[Shi 2004] C. Shi and R. Kapur, "How Power Aware Test Improves Reliability and Yield," *IEEDesign.com*, September 15, 2004.

[Ciri87] M.A. Cirit, "Estimating Dynamic Power Consumption of CMOS Circuits," *Proc. Int'l Conf. on Computer-Aided Design*, pp. 534-537, November 1987.

[Wang 1995] C.-Y. Wang and K. Roy, "Maximum Power Estimation for CMOS Circuits Using Deterministic and Statistical Approaches," *Proc. VLSI Conf.*, pp. 364-369, January 1995.

R7.3 Test Power Issues

[Butler 2004] K. M. Butler, J. Saxena, T. Fryars, G. Hetherington, A. Jain, and J. Lewis, "Minimizing Power Consumption in Scan Testing: Pattern Generation and DFT Techniques," *Proc. Int'l Test Conf.*, pp. 355-364, October 2004.

[Chang 1997] Y.-S. Chang, S. K. Gupta and M. A. Breuer, "Analysis of Ground Bounce in Deep Sub-Micron Circuits," *Proc. VLSI Test Symp.*, pp. 110-116, May 1997.

[Hertwig 1998] A. Hertwig and H.J. Wunderlich, "Low Power Serial Built-In Self-Test," *Proc. European Test Workshop*, pp. 49-53, May 1998.

[Jiang 2000] Y.M. Jiang, A. Krstic and K.-T. Cheng, "Estimation for Maximum Instantaneous Current Through Supply Lines for CMOS Circuits," *IEEE Trans. on Very Large Scale Integration Systems*, Vol. 8, No. 1, pp. 61-73, Feb. 2000.

[Nicolaidis 1998] M. Nicolaidis and Y. Zorian, "On-Line Testing for VLSI: A Compendium of Approaches," *JETTA Journal of Electronic Testing – Theory and Applications*, Vol. 12, No. 1-2, pp. 7-20, Feb./Apr. 1998.

[Pouya 2000] B. Pouya and A. Crouch, "Optimization Trade-offs for Vector Volume and Test Power," *Proc. Int'l Test Conf.*, pp. 873-881, October 2000.

[SIA 2003] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors: 2003 Edition*, http://public.itrs.net, 2003.

[Shi 2004] C. Shi and R. Kapur, "How Power Aware Test Improves Reliability and Yield," *IEEDesign.com*, September 15, 2004.

[Wang 1997] S. Wang and S.K. Gupta, "DS-LFSR: A New BIST TPG for Low Heat Dissipation," *Proc. Int'l Test Conf.*, pp. 848-857, November 1997.

R7.4 Low-Power Scan Testing

[Badereddine 2006] N. Badereddine, P. Girard, S. Pravossoudovitch, C. Landrault, A. Virazel, and H. J. Wunderlich, "Minimizing Peak Power Consumption during Scan Testing: Test Pattern Modification with X Filling Heuristics," *Proc. Int'l Conf. on Design & Test of Integrated Systems*, pp. 259-264, September 2006.

[Bonhomme 2001] Y. Bonhomme, P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "A Gated Clock Scheme for Low Power Scan Testing of Logic ICs or Embedded Cores," *Proc. Asian Test Symp.*, pp. 253-258, November 2001.

[Bonhomme 2002] Y. Bonhomme, P. Girard, C. Landrault, and S. Pravossoudovitch, "Power Driven Chaining of Flip-flops in Scan Architectures," *Proc. Int'l Test Conf.*, pp. 796-803, October 2002.

[Bonhomme 2003] Y. Bonhomme, P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, “Efficient Scan Chain Design for Power Minimization During Scan Testing Under Routing Constraint,” *Proc. Int’l Test Conf.*, pp. 488-493, October 2003.

[Butler 2004] K. M. Butler, J. Saxena, T. Fryars, G. Hetherington, A. Jain, and J. Lewis, “Minimizing Power Consumption in Scan Testing: Pattern Generation and DFT Techniques,” *Proc. Int’l Test Conf.*, pp. 355-364, October 2004.

[Dabholkar 1998] V. Dabholkar, S. Chakravarty, I. Pomeranz, and S. M. Reddy, “Techniques for Reducing Power Dissipation During Test Application in Full Scan Circuits,” *IEEE Trans. on Computer-Aided Design*, Vol. 17, No. 12, pp. 1325-1333, Dec. 1998.

[Girard 2002] P. Girard, “Survey of Low-Power Testing of VLSI Circuits,” *IEEE Design & Test of Computers*, Vol. 19, No. 3, pp. 82-92, May-June 2002.

[Hiraide 2003] T. Hiraide, K. O. Boateng, H. Konishi, K. Itaya, M. Emori, H. Yamanaka, and T. Mochiyama, “BIST-Aided scan test – a new method for test cost reduction,” *Proc. VLSI Test Symp.*, pp. 359-364, May 2003.

[Hertwig 1998] A. Hertwig and H. J. Wunderlich, “Low Power Serial Built-In Self-Test,” *Proc. European Test Workshop*, pp. 49-53, May 1998.

[Huang 1999] T.-C. Huang and K.-J. Lee, “An Input Control Technique for Power Reduction in Scan Circuits During Test Application,” *Proc. Asian Test Symp.*, pp. 315-320, November 1999.

[Huang 2001] T.-C. Huang and K.-J. Lee, “A Token Scan Architecture for Low Power Testing,” *Proc. Int’l Test Conf.*, pp. 660-669, October 2001.

[Lee 2000] K.-J. Lee, T.-C. Huang, and J.-J. Chen, “Peak-Power Reduction for Multiple-Scan Circuits during Test Application,” *Proc. IEEE Asian Test Symp.*, pp. 453-458, December 2000.

[Remersaro 2006] S. Remersaro, X. Lin, Z. Zhang, S. M. Reddy, I. Pomeranz, and J. Rajski, “Preferred Fill: A Scalable Method to Reduce Capture Power for Scan Based Designs,” *Proc. Int’l Test Conf.*, Paper 32.2, October 2006.

[Rosinger 2004] P. Rosinger, B. Al-Hashimi, and N. Nicolici, “Scan Architecture with Mutually Exclusive Scan Segment Activation for Shift- and Capture-Power Reduction,” *IEEE Trans. on Computer-Aided Design*, Vol. 23, No. 7, pp. 1142-1153, July 2004.

[Sankaralingam 2000] R. Sankaralingam, R. Oruganti, and N. A. Touba, “Static Compaction Techniques to Control Scan Vector Power Dissipation,” *Proc. VLSI Test Symp.*, pp. 35-42 , May 2000.

[Sankaralingam 2003] R. Sankaralingam and N. A. Touba, “Multi-Phase Shifting to Reducing Instantaneous Peak Power During Scan,” *Proc. Latin American Test Workshop*, pp. 78-83, February 2003.

[Saxena 2001] J. Saxena, K. M. Butler, and L. Whetsel, "A Scheme to Reduce Power Consumption During Scan Testing," *Proc. Int'l Test Conf.*, pp. 670-677, October 2001.

[Saxena 2003] J. Saxena, K. M. Butler, V. B. Jayaram, S. Kundu, N. V. Arvind, P. Sreeprakash, and M. Hachinger, "A Case Study of IR-Drop in Structured At-Speed Testing," *Proc. Int'l Test Conf.*, pp. 1098-1104, October 2003.

[Shi 2004] C. Shi and R. Kapur, "How Power Aware Test Improves Reliability and Yield," *IEEDesign.com*, September 15, 2004.

[Sinanoglu 2002] O. Sinanoglu, I. Bayraktaroglu, and A. Orailoglu, "Dynamic Test Data Transformations for Average and Peak Power Reductions," *Proc. European Test Workshop*, pp. 113-118, May 2002.

[Wang 1994] S. Wang and S. K. Gupta, "ATPG for Heat Dissipation Minimization During Test Application," *Proc. Int'l Test Conf.*, pp. 250-258, October 1994.

[Wang 1997] S. Wang and S. K. Gupta, "ATPG for Heat Dissipation Minimization for Scan Testing," *Proc. Design Automation Conf.*, pp. 614-619, June 1997.

[Wen 2005a] X. Wen, Y. Yamashita, S. Morishima, S. Kajihara, L.-T. Wang, K. K. Saluja, and K. Kinoshita, "Low-Capture-Power Test Generation for Scan-Based At-Speed Testing," *Proc. Int'l Test Conf.*, Paper 39.2, November 2005.

[Wen 2005b] X. Wen, T. Suzuki, S. Kajihara, K. Miyase, Y. Minamoto, L.-T. Wang, and K.K. Saluja, "Efficient Test Set Modification for Capture Power Reduction," *ASP Journal of Low Power Electronics*, Vol. 1, No. 3, pp. 319-330, Dec. 2005.

[Wen 2006] X. Wen, S. Kajihara, K. Miyase, T. Suzuki, K. K. Saluja, L.-T. Wang, K. S. Abdel-Hafez, and K. Kinoshita, "A New ATPG Method for Efficient Capture Power Reduction During Scan Testing," *Proc. VLSI Test Symp.*, pp. 58-63, May 2006.

[Whetsel 2000] L. Whetsel, "Adapting Scan Architectures for Low Power Operation," *Proc. Int'l Test Conf.*, pp. 863-872, October 2000.

[Wohl 2003] P. Wohl, J. A. Waicukauski, S. Patel, and M. B. Amin, "Efficient Compression and Application of Deterministic Patterns in a Logic BIST Architecture," *Proc. Design Automation Conf.*, pp. 566-569, June 2003.

[Xu 2006] G. Xu and A. D. Singh, "Low Cost Launch-on-Shift Delay Test with Slow Scan Enable," *Proc. European Test Symp.*, Paper 3a-1, May 2006.

R7.5 Low-Power Built-In Self-Test

[Brglez 1985] F. Brglez and H. Fujiwara, "A Neutral Netlist of 10 Combinational Benchmark Circuits and a Target Translator in Fortran," *Proc. Int'l Symp. on Circuits and Systems*, pp. 663-698, June 1985.

[Chou 1994] R.-M. Chou, K. K. Saluja and V. D. Agrawal, "Power Constraint Scheduling of Tests," *Proc. Int'l Conf. on VLSI Design*, pp. 271-274, January 1994.

[Corno 1999] F. Corno, M. Rebaudengo, M. Sonza Reorda, and M. Violante, "A New BIST Architecture for Low Power Circuits," *Proc. European Test Workshop*, pp. 160–164, May 1999.

[Corno 2000] F. Corno, M. Rebaudengo, M. Sonza Reorda, G. Squillero, and M. Violente, "Low Power BIST via Non-Linear Hybrid Cellular Automata," *Proc. VLSI Test Symp.*, pp. 29-34, May 2000.

[Gerstendörfer 1999] S. Gerstendörfer and H. J. Wunderlich, "Minimized Power Consumption for Scan-Based BIST," *Proc. Int'l Test Conf.*, pp. 77-84, September 1999.

[Girard 1999a] P. Girard, L. Guiller, C. Landrault, S. Pravossoudovitch, J. Figueras, S. Manich, P. Teixeira, and M. Santos, "Low Energy BIST Design: Impact of the LFSR TPG Parameters on the Weighted Switching Activity," *Proc. Int'l Symp. on Circuits and Systems*, CD-ROM Proceedings, June 1999.

[Girard 1999b] P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "A Test Vector Inhibiting Technique for Low Energy BIST Design," *Proc. VLSI Test Symp.*, pp. 407-412, April 1999.

[Girard 1999c] P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "Circuit Partitioning for Low Power BIST Design with Minimized Peak Power Consumption," *Proc. Asian Test Symp.*, pp. 89-94, November 1999.

[Girard 2000] P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "Low Power BIST Design by Hypergraph Partitioning: Methodology and Architectures," *Proc. Int'l Test Conf.*, pp. 652-661, October 2000.

[Girard 2001] P. Girard, L. Guiller, C. Landrault, S. Pravossoudovitch, and H. J. Wunderlich, "A Modified Clock Scheme for a Low Power BIST Test Pattern Generator," *Proc. VLSI Test Symp.*, pp. 306-311, May 2001.

[Gizopoulos 2000] D. Gizopoulos, N. Kranitis, A. Paschalis, M. Psarakis, and Y. Zorian, "Low Power/Energy BIST Scheme for Datapaths," *Proc. VLSI Test Symp.*, pp. 23-28, May 2000.

[He 2006] Z. He, Z. Peng, P. Eles, P. Rosinger, and B. Al-Hashimi, "Thermal-Aware SOC Test Scheduling with Test Set Partitioning and Interleaving," *Proc. Int'l Symp. on Defect and Fault Tolerance in VLSI Systems*, pp. 477-485, October 2006.

[Iyengar 2001] V. Iyengar and K. Chakrabarty, "Precedence-Based, Preemptive, and Power-Constrained Test Scheduling for System-on-a-Chip," *Proc. VLSI Test Symp.*, pp. 42-47, May 2001.

[Karypis 1998] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel Hypergraph Partitioning: Applications in VLSI Domain," Technical Report, Department of Computer Science, University of Minnesota, November 1998. http://www.cs.umn.edu/~karypis/metis.

[Larsson 2002] E. Larsson and H. Fujiwara, "Power-Constrained Preemptive TAM Scheduling," *Proc. European Test Workshop*, pp. 119-126, May 2002.

[Lee 2004] J. Lee and N. A. Touba, "Low Power Test Data Compression Based on LFSR Reseeding," *Proc. Int'l Conf. on Computer Design*, pp. 180-185, October 2004.

[Liu 2005] C. Liu, K. Veeraraghavant, and V. Iyengar, "Thermal-Aware Test Scheduling and Hot Spot Temperature Minimization for Core-Based Systems," *Proc. Int'l Symp. on Defect and Fault Tolerance in VLSI Systems*, pp. 552-562, October 2005.

[Manich 2000] S. Manich, A. Gabarro, M. Lopez, J. Figueras, P. Girard, L. Guiller, C. Landrault, S. Pravossoudovitch, P. Teixeira, and M. Santos, "Low Power BIST by Filtering Non-Detecting Vectors," *JETTA Journal of Electronic Testing – Theory and Applications*, Vol. 16, No. 3, pp. 193-202, June 2000.

[Muresan 2000] V. Muresan, X. Wang, and M. Vladutiu, "A Comparison of Classical Scheduling Approaches in Power-Constrained Block-Test Scheduling," *Proc. Int'l Test Conf.*, pp. 882-891, October 2000.

[Pouget 2003]J. Pouget, E. Larsson, Z. Peng, M.L. Flottes, and B. Rouzeyre, "An Efficient Approach to SOC Wrapper Design, TAM Configuration and Test Scheduling," *Proc. European Test Workshop*, pp. 117-122, May 2003.

[Rosinger 2005] P. Rosinger, B. Al-Hashimi, and K. Chakrabarty, "Rapid Generation of Thermal-Safe Test Schedules," *Proc. Design, Automation and Test in Europe*, pp. 840-845, March 2005.

[Ubar 2005] R. Ubar, T. Shchenova, G. Jervan, and Z. Peng, "Energy Minimization for Hybrid BIST in a System-on-Chip Test Environment," *Proc. European Test Symp.*, pp. 2-7, May 2005.

[Wang 1997] S. Wang and S. K. Gupta, "DS-LFSR: A New BIST TPG for Low Heat Dissipation," *Proc. Int'l Test Conf.*, pp. 848-857, November 1997.

[Wang 1999] S. Wang and S. K. Gupta, "LT-RTPG: A New Test-Per-Scan BIST TPG for Low Heat Dissipation," *Proc. Int'l Test Conf.*, pp. 85-94, September 1999.

[Zhang 1999] X. Zhang, K. Roy, and S. Bhawmik, "POWERTEST: A Tool for Energy Concious Weighted Random Pattern Testing," *Proc. Int'l Conf. on VLSI Design*, pp. 416-422, January 1999.

[Zorian 1993] Y. Zorian, "Testing the Monster Chip," *IEEE Spectrum*, Vol. 36, No. 7, pp. 54-60, July 1999.

R7.6 Low-Power Test Data Compression

[Al-Yamani 2005] A. Al-Yamani, E. Chmelar, and M. Grinchuck, “Segmented Addressable Scan Architecture,” *Proc. VLSI Test Symp.*, pp. 405-411, May 2005.

[Baik 2005] D.H. Baik and K.K. Saluja, “Progressive Random Access Scan: A Simultaneous Solution to Test Power, Test Data Volume and Test Time,” *Proc. Int’l Test Conf.*, paper 15.2, November 2005.

[Chandra 2001] A. Chandra and K. Chakrabarty, “Combining Low-Power Scan Testing and Test Data Compression for System-on-a-Chip,” *Proc. Design Automation Conf.*, pp. 166-169, June 2001.

[Chandra 2002] A. Chandra and K. Chakrabarty, “Reduction of SOC Test Data Volume, Scan Power and Testing Time Using Alternating Run-Length Codes,” *Proc. Design Automation Conf.*, pp. 673-678, June 2002.

[Hamzaoglu 1999] I. Hamzaoglu and J. Patel, “Reducing Test Application Time for Full Scan Embedded Cores,” *Proc. Int’l Symp. on Fault Tolerant Computing*, pp. 206-267, 1999.

[Lee 2004] J. Lee and N. A. Touba, “Low Power Test Data Compression Based on LFSR Reseeding,” *Proc. Int’l Conf. on Computer Design*, pp. 180-185, October 2004.

[Pouya 2000] B. Pouya and A. Crouch, “Optimization Trade-offs for Vector Volume and Test Power,” *Proc. Int’l Test Conf.*, pp. 873-881, October 2000.

R7.7 Low-Power RAM Testing

[Cheung 1996] H. Cheung and S. Gupta, “A BIST Methodology for Comprehensive Testing of RAM with Reduced Heat Dissipation,” *Proc. Int’l Test Conf.*, pp. 22-32, October 1996.

[Dilillo 2006] L. Dilillo, P. Rosinger, P. Girard, and B. M. Al-Hashimi, “Minimizing Test Power in SRAM through Pre-charge Activity Reduction,” *Proc. Design, Automation and Test in Europe*, pp. 1159-1165, March 2006.

[Liu 1994] D. Liu and C. Svensson, “Power Consumption Estimation in CMOS VLSI Chips,” *IEEE Journal of Solid-State Circuits*, Vol. 29, No. 6, pp. 663-670, June 1994.

R7.8 Summary

[SIA 2005] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors: 2005 Edition*, http://public.itrs.net, 2005.

[SIA 2006] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors: 2006 Update*, http://public.itrs.net, 2006.