

Chapter 2

Power Issues During Test

Sandip Kundu and Alodeep Sanyal

Abstract An unintended consequence of technology scaling has increased power consumption in a chip. Without specialized solutions, level of power consumption and rate of change of power consumption is even greater during test. Power delivery during test is somewhat limited by mechanical and electrical constraints. This chapter introduces the basic concepts related to power and energy and describes typical manufacturing test flow and associated constraints with power delivery. It also describes various types of power droop mechanisms, thermal issues, and how they interfere with the test process. Test economics issues, such as throughput and yield loss, are also discussed to further develop the low-power test problem statement.

2.1 Introduction

Continuous scaling of the feature size of complementary metal oxide semiconductor (CMOS) technology has resulted in exponential growth in transistor densities, enabling more functionality to be placed on a silicon die. The growth in transistor density has been accompanied with linear reduction in the supply voltage that has not been adequate in keeping power densities from rising. Elevated power densities lead to a two pronged problem: (1) supplying adequate power for circuit operation and (2) a heat flux from resulting dissipation. The power delivery issue can lead to supply integrity problems, whereas the heat flux issue affects packaging at chip, module, and system levels. In several situations, the form factor dictates a thermal envelope. Many modern systems from mobile to high-performance computers implement power management to address both energy and thermal envelope issues (Nicolici and Wen 2007).

S. Kundu (✉) and A. Sanyal
University of Massachusetts, Amherst, MA, USA
e-mail: kundu@ecs.umass.edu

Power issues are not confined to functional operation of devices only. They also manifest during testing. First, power consumption may rise during testing (Dabholkar et al. 1998, Girard 2002, Nicolici and Wen 2007):

- Typical power management schemes are disabled during testing leading to increased power consumption.
 - Clock gating is turned off to improve observability of internal nodes during testing.
 - Dynamic frequency scaling is turned off during test either because the system clock is bypassed or because the *phase locked loop* (PLL) suffers from a relocking time overhead during which no meaningful test can be conducted.
 - Dynamic voltage scaling is usually avoided due to time constants in stabilizing supply voltage.
- Switching activity may be higher during testing.
 - Because of *automatic test pattern generation* (ATPG) complexity, testing is predominantly done structurally. Structural testing tends to produce more toggling than functional patterns because the goal of (structural) testing is to activate as many nodes as possible in the shortest test time, which is not the case during functional mode. Another reason is that the design-for-testability (DFT) (e.g., scan) circuitry is intensively used and stresses the circuit-under-test (CUT) much more than during functional mode.
 - Test compaction leads to higher switching activity due to parallel fault activation and propagation in a circuit.
 - Multiple cores in a *system-on-a-chip* (SoC) are tested in parallel to reduce test application time, which inherently lead to significant rise in switching activity.

Second, power availability and quality may be limited during testing:

- Longer connectors from *tester power supply* (TPS) to probe-card often result in higher inductance on the power delivery path. This may lead to voltage drop during test power cycling.
- During wafer sort test, all power pins may not be connected to the TPS, resulting in reduced power availability.
- Current limiters placed on TPS to prevent burn-out due to short-circuit current may interfere with both availability and quality of supply voltage during power surges that may result from testing.
- Reduced power availability may impact performance and in some cases may lead to loss of correct logic state of the device resulting in *manufacturing yield loss*.

Finally, there may be a reliability aspect of power to be considered during testing:

- Bus contention problem: during structural testing, nonfunctional vectors may cause illegal circuit operation such as creating a path from V_{DD} to ground with short circuit power dissipation.

- Memory contention problem: this occurs in a multiported memory, where simultaneous writes with conflicting data may take place to the same address, typically by nonfunctional patterns applied during structural testing.
- Bus and memory contention problems may cause short-circuit and permanent damage to the device. Therefore, it is important to conduct electrical verification of test vectors from a circuit operation point of view before they are applied from a tester.

In the rest of the chapter, we will explore these issues in greater depth. Subsequent sections in this chapter provide introduction to basic concepts related to power and energy, describe typical manufacturing test flow with staggered, multifaceted test objectives to provide a context for power and thermal issues during test. The discussion of test issues regarding power is contextualized with constraints arising out of test instrument, environment, test patterns, and test economics.

2.2 Power and Energy Basics

There are two major components of power dissipation in a CMOS circuit (Rabaey et al. 1996, Weste and Eshraghian 1988):

- Static dissipation due to leakage current or other currents drawn continuously from the power supply.
- Dynamic dissipation due to
 - charging and discharging of load capacitances and
 - short-circuit current.

The following two subsections discuss briefly the individual power components with the aid of a CMOS inverter.

2.2.1 Static Dissipation

The static (or steady-state) power dissipation of a circuit is given by the following expression:

$$P_{\text{stat}} = \sum_{i=1}^n I_{\text{stat}_i} \cdot V_{\text{DD}} \quad (2.1)$$

where I_{stat} is the current that flows between the supply rails in the absence of switching activity and i is the index of a gate in a circuit consisting of n gates.

Ideally, the static current of the CMOS inverter is equal to zero, as the positive and negative metal oxide semiconductor (PMOS and NMOS) devices are never ON simultaneously in the steady-state operation. However, there are some leakage currents that cause static power dissipation. The sources of leakage current

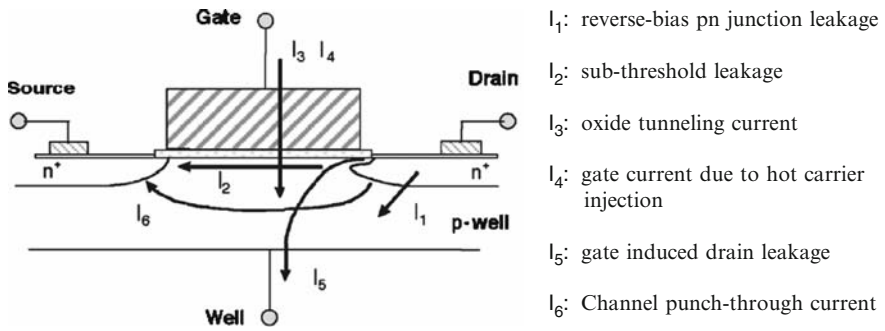


Fig. 2.1 Summary of leakage current mechanisms of deep-submicron transistors [figure adopted from Roy et al. (2003)]

for a CMOS inverter are indicated in Fig. 2.1. Major leakage contributors are (1) reverse-biased leakage current (I_1) between source and drain diffusion regions and the substrate, (2) sub-threshold conduction current (I_2) between source and drain, and (3) pattern-dependent leakage (I_3) across gate oxide (Roy et al. 2003). The other two sources of leakage which are often taken into consideration (especially as we move into nanometer design) are (1) gate-induced drain leakage (GIDL) and (2) drain-induced barrier lowering. In the following text, we describe them with some details.

2.2.1.1 Reverse-Biased pn Junction Leakage Current

Drain and source to well junctions are typically reverse-biased, causing pn junction leakage current (I_1). A reverse-biased pn junction leakage has two main components: (1) minority carrier diffusion/drift near the edge of the depletion region and (2) electron-hole pair generation in the depletion region of the reverse-biased junction (Pierret 1996). If both n and p regions are heavily doped as is the case for nanoscale CMOS devices, the depletion width is smaller and the electric field across depletion region is higher. Under this condition ($E > 1 \text{ MV/cm}$), direct band-to-band tunneling (BTBT) of electrons from the valence band of the p region to the conduction band of the n region becomes significant. In nanoscale CMOS circuits, BTBT leakage current dominates the pn junction leakage. For tunneling to occur, the total voltage drop across the junction has to be more than the band gap (Roy et al. 2003). Logical bias conditions for I_{BTBT} are shown in Fig. 2.2.

2.2.1.2 Sub-threshold Leakage Current

Sub-threshold current is the most dominant among all sources of leakage. It is caused by minority carriers drifting across the channel from drain to source due to the presence of weak inversion layer when the transistor is operating in cut-off

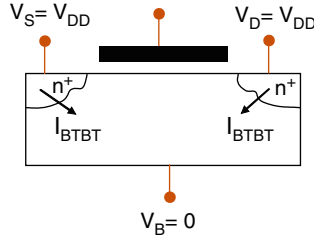


Fig. 2.2 Illustration of band-to-band tunneling (BTBT) leakage in a negative metal oxide semiconductor (NMOS) transistor

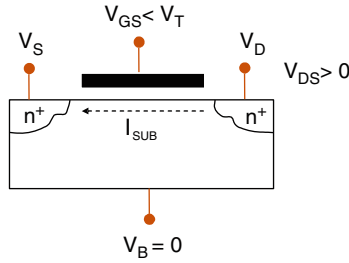


Fig. 2.3 Illustration of sub-threshold leakage in a negative metal oxide semiconductor (NMOS) transistor

region ($V_{GS} < V_t$). The minority carrier concentration rises exponentially with gate voltage V_G . The plot of $\log(I_2)$ versus V_G is a linear curve with typical slopes of 60–80 mV per decade. Sub-threshold leakage current depends on the channel doping concentration, channel length, threshold voltage V_t , and the temperature. In Fig. 2.3, the bias condition for sub-threshold current (I_{SUB}) on an NMOS device has been illustrated.

2.2.1.3 Gate Leakage Current

Reduction of gate oxide thickness results in an increase in the electric field across the oxide. The high electric field coupled with low oxide thickness results in tunneling of electrons from substrate to gate and also from gate-to-substrate through the gate oxide, resulting in the gate oxide tunneling current. The mechanism of tunneling between substrate and gate polysilicon can be primarily divided into two parts, namely: (1) *Fowler-Nordheim tunneling* and (2) *direct tunneling*. In the case of *Fowler-Nordheim* tunneling, electrons tunnel through a triangular potential barrier, whereas in the case of *direct tunneling*, electrons tunnel through a trapezoidal potential barrier. The tunneling probability of an electron depends on the thickness of the barrier, the barrier height, and the structure of the barrier (Roy et al. 2003).

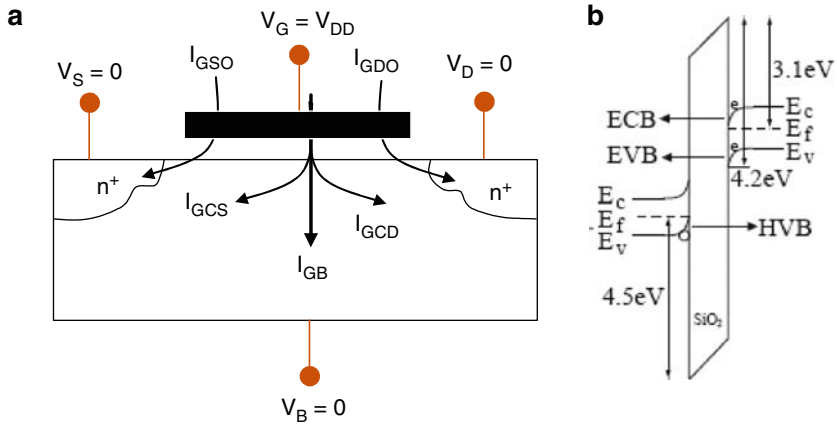


Fig. 2.4 (a) Illustration of gate leakage in a negative metal oxide semiconductor (NMOS) device and (b) the tunneling mechanism in band diagram [figure adopted from Drazdziulis and Larsson-Edefors (2003)]

The gate tunneling current can be divided into five major components, namely, parasitic leakage current through gate-to-source/drain extension overlap region (I_{GSO} and I_{GDO}); gate-to-inverted channel current (I_{GC}), part of which goes to the source (I_{GCS}) and the rest goes to the drain (I_{GCD}) (Hu et al. 2000); and the gate-to-substrate leakage current (I_{GB}). I_{GSO} and I_{GDO} are parasitic leakage currents that pass through gate-to-source/drain extension overlap region. I_{GDO} in off-state ($V_G = 0$) NMOS device is also known as *edge directed tunneling current* (EDL) (Yang et al. 2001) and is higher than its on-state counterpart. PMOS devices have less gate leakage compared with NMOS devices as holes have higher barrier of 4.5 eV compared with 3.1 eV for electron. Total gate leakage current is given as:

$$I_2 = I_{GSO} + I_{GDO} + I_{GCS} + I_{GCD} + I_{GB} \quad (2.2)$$

A bias condition at which I_{GC} , I_{GB} , and I_{EDL} occur is shown in Fig. 2.4.

2.2.1.4 Gate-Induced Drain Leakage Current

GIDL is due to high field effect in the drain junction of an MOS transistor. When the gate is biased to form an accumulation layer at the silicon surface, the silicon surface under the gate has almost same potential as the p-type substrate. Because of the presence of accumulated holes at the surface, the surface behaves like a p region more heavily doped than the substrate. This causes the depletion layer at the surface to be much narrower than elsewhere. The narrowing of the depletion layer at or near the surface causes field crowding or an increase in the local electric field, thereby enhancing the high field effects near that region. Large negative gate bias increases

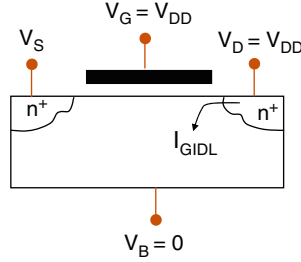


Fig. 2.5 Illustration of gate-induced drain leakage (*GIDL*) leakage in a negative metal oxide semiconductor (*NMOS*) transistor

field crowding further and peak field also increases, and the possibility of tunneling via near-surface traps also increases (Taur and Ning 1998). As a result of all these effects, minority carriers are emitted in the drain region underneath the gate. Since the substrate is at a lower potential for minority carriers, the minority carriers that have been accumulated or formed at the drain depletion region underneath the gate are swept laterally to the substrate, completing a path for the GIDL (Roy et al. 2003).

GIDL current is gaining importance as we move deeper into nanometer technologies. In Fig. 2.5, I_{GIDL} is illustrated for an NMOS device. Explanation of I_{GIDL} in PMOS can be similarly described.

2.2.2 Dynamic Dissipation

2.2.2.1 Dynamic Dissipation Due to Charging and Discharging of Load Capacitors

For a CMOS inverter, the dynamic power is dissipated mainly due to charging and discharging of the load capacitance (lumped as C_L as shown in Fig. 2.6). When the input to the inverter is switched to logic state 0 (Fig. 2.6a), the PMOS is turned ON and the NMOS is turned OFF. This establishes a resistive DC path from power supply rail to the inverter output and the load capacitor C_L starts charging, whereas the inverter output voltage rises from 0 to V_{DD} . During this charging phase, a certain amount of energy is drawn from the power supply. Part of this energy is dissipated in the PMOS device which acts as a resistor, whereas the remainder is stored on the load capacitor C_L . During the high-to-low transition (Fig. 2.6b), the NMOS is turned ON and the PMOS is turned OFF, which establishes a resistive DC path from the inverter output to the Ground rail. During this phase, the capacitor C_L is discharged, and the stored energy is dissipated in the NMOS transistor (Cirit 1987, Rabaey et al. 1996, Weste and Eshraghian 1988).

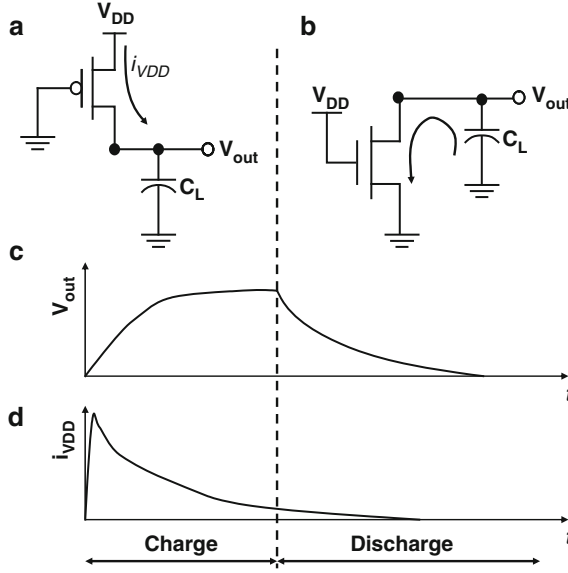


Fig. 2.6 Equivalent circuit during the (a) low-to-high transition, (b) high-to-low transition, (c) output voltages, and (d) supply current during corresponding charging and discharging phases of C_L [figure adopted from [Rabaey et al. \(1996\)](#)]

A precise measure for this energy consumption can be derived. Let us first consider the low-to-high transition. We start with a simplifying assumption that the NMOS and PMOS devices have zero *rise* and *fall times*, or in other words, the NMOS and PMOS devices are never ON simultaneously. Under this assumption, the equivalent circuits for charging and discharging of the load capacitor as shown in Fig. 2.6a,b are valid. The expressions for the energy E_{VDD} , taken from the supply during the transition, as well as the energy E_C , stored on the load capacitor at the end of the transition, can be derived by integrating the instantaneous power over the period of interest ([Rabaey et al. 1996](#)):

$$E_{VDD} = \int_0^\infty i_{VDD}(t) V_{DD} dt = V_{DD} \int_0^\infty C_L \frac{dv_{out}}{dt} dt = C_L V_{DD} \int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2 \quad (2.3)$$

and

$$E_C = \int_0^\infty i_{VDD}(t) v_{out} dt = \int_0^\infty C_L \frac{dv_{out}}{dt} v_{out} dt = C_L \int_0^{V_{DD}} v_{out} dv_{out} = \frac{1}{2} C_L V_{DD}^2 \quad (2.4)$$

The corresponding waveforms of $v_{out}(t)$ and $i_{VDD}(t)$ are depicted in Fig. 2.6c,d, respectively. From (2.3) and (2.4), we may infer that only half of the energy supplied by the power source is stored on C_L . The other half is dissipated as heat by the

PMOS transistor that acts as a resistor. During the discharge phase, the charge is removed from the load capacitor, and its energy gets dissipated as heat in the NMOS transistor forming a resistive path to the Ground. In summary, each switching cycle (consisting of an $L \rightarrow H$ and an $H \rightarrow L$ transition) takes a fixed amount of energy, equal to $C_L V_{DD}^2$. In order to compute the power consumption, we have to take into account how often the device is switched. If the inverter is switched on and off during a given time period, the power consumption is given by

$$P_d = C_L V_{DD}^2 f_{0 \rightarrow 1} \quad (2.5)$$

where $f_{0 \rightarrow 1}$ represents the number of rising transitions at the inverter output per second.

2.2.2.2 Dynamic Dissipation Due to Short-Circuit Current

Even though under the simplifying assumption of zero rise and fall times for NMOS and PMOS devices for static CMOS logic gates, there exists no direct current path between the power and ground rails, a more realistic timing model for CMOS technology reveals that the input switching is gradual and not abrupt. Consequently, during switching of input, the PMOS and NMOS devices remain ON simultaneously for a finite period. The current associated with this DC current between supply rails is known as short-circuit current (I_{sc}) (Veendrick 1984, Vemuri and Scheinberg 1994, Hirata et al. 1996). Since short-circuit power is delivered by the voltage supply (V_{DD}), the total power can be written as

$$P_{sc} = V_{DD} \int_T I_{sc}(\tau) d\tau \quad (2.6)$$

where T is the switching period (Acar et al. 2003).

Let us now analyze the short-circuit power component with the aid of a rising ramp input applied to a CMOS inverter as shown in Fig. 2.7. Assuming the input signal begins to rise at origin, the time interval for short-circuit current starts at t_0 when the NMOS device turns ON, and ends at t_1 when the PMOS device turns OFF. During this time interval, the PMOS device moves from linear region of operation

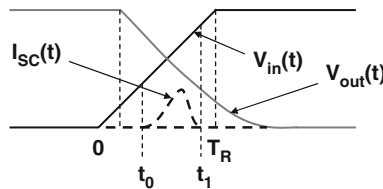


Fig. 2.7 Input and output waveforms for a complimentary metal oxide semiconductor (CMOS) inverter when the input switches from low to high and the corresponding short circuit current [figure adopted from Acar et al. (2003)]

to saturation region. On the basis of the ramp input signal with a rise time T_R (as shown in Fig. 2.7), t_0 and t_1 can be expressed as:

$$t_0 = T_R \frac{V_{thn}}{V_{DD}} \quad (2.7a)$$

$$t_1 = T_R \frac{V_{DD} + V_{thp}}{V_{DD}} \quad (2.7b)$$

The average short-circuit power can be specified as the integral of short-circuit current between t_0 and t_1 :

$$P_{sc} = V_{DD} \int_{t_0}^{t_1} \frac{I_{sc}(\tau) d\tau}{(t_1 - t_0)} \quad (2.8)$$

2.2.3 Total Power Dissipation

The total power consumption of the CMOS inverter is now expressed as the sum of its three components:

$$P_{total} = P_{stat} + P_d + P_{sc} \quad (2.9)$$

In typical CMOS circuits, the capacitive dissipation was by far the dominant factor. However, with the advent of deep-submicron regime in CMOS technology, the static (or leakage) consumption of power has grown rapidly and account for more than 25% of power consumption in SoCs and 40% of power consumption in high performance logic (ITRS 2007).

2.2.4 Energy Dissipation

Energy is defined as the total power consumed in a CMOS circuit over a period of T . Therefore, mathematically we may express energy dissipated in a CMOS circuit as:

$$E_{total} = \int_T P_{total} d\tau \quad (2.10)$$

Substituting the expression for P_{total} from (2.9), we get:

$$E_{total} = \int_T P_{stat} d\tau + \int_T P_d d\tau + \int_T P_{sc} d\tau \quad (2.11)$$

All the three individual power components are input state dependent. Therefore, the energy dissipated over a period of T will depend on the set of input vectors applied to the circuit during that period as well as the order in which they are applied.

2.3 Manufacturing Test Flow

Testing and diagnosis of VLSI systems can be broadly classified into four types depending on the specific purpose it accomplishes and the current phase of production (from fabrication to shipment) for the circuit under test (Bushnell and Agrawal 2000, Stevens 1986). In the following four subsections, we briefly cover these four types of test methods in the order they are conducted during the design and manufacturing processes.

2.3.1 Characterization Test

Also known as design debug or verification testing, this form of testing is performed on a new design before it is sent to production (Bushnell and Agrawal 2000). The main objective of characterization test is to verify that the design is correct and the device will meet all specifications. Comprehensive AC and DC measurements are made during this test process. The requirement for thoroughness during this testing phase may often lead to probing of internal nodes of a chip, not performed as part of any other test process. Specialized tools such as *scanning electron microscopes* and electron beam testers, and techniques such as *artificial intelligence* and *expert systems* are often used in this form of testing. A characterization test determines the exact limits of device operating values. Generally, the devices are tested for the worst case because it is easier to evaluate than average cases and devices passing this test will work for any other conditions.

2.3.2 Production Test

Every fabricated chip is subjected to production tests, which are less comprehensive than characterization tests yet they must enforce the quality requirements by determining whether the device meets specifications (Bushnell and Agrawal 2000). It may not be possible to cover all possible functions and data patterns, but production tests must have a high coverage of modeled faults. Since every device must be tested before being packaged, test application time is of great importance. Production test should be as brief as possible and is usually different from characterization tests or diagnostic tests.

2.3.3 Burn-in Test

All devices that pass production tests are not identical. When put to actual use, some will fail very quickly whereas others will function for a long time. Burn in screens for long-term reliability of devices by either continuous or periodic testing over a

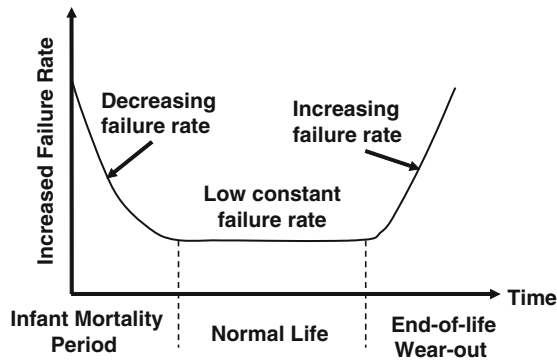


Fig. 2.8 Bathtub curve showing the rate of failure of integrated circuits at different phases of life [figure adopted from [Hnatek \(1987\)](#)]

period, usually under nonrated conditions. Rate of failure of integrated circuits at different phases of life follows a *bathtub* curve ([Hnatek 1987](#)) (shown in Fig. 2.8). Correlation studies show that the occurrence of potential failures can be accelerated at elevated temperatures ([Jensen and Petersen 1982](#)). Two types of failures are isolated by burn-in: (1) *infant mortality failures*, often caused by a combination of sensitive design and process variation, and may be screened out by a short-term burn-in (10–30 h) in a normal or slightly accelerated working environment, and (2) *freak failures*, that is, devices having the same failure mechanisms as the reliable devices, require long burn-in time (100–1,000 h) in an accelerated environment. In practice, a manufacturer must balance economic considerations against the device reliability. In any case, the elimination of infant mortality failures is considered essential in many applications ([Bushnell and Agrawal 2000](#)).

2.3.4 Incoming Inspection

System manufacturers perform incoming inspection (also called *quality assurance*) on the purchased devices before integrating them into the system. Depending upon the context, this testing can be either similar to production testing, or more comprehensive than production testing, or even tuned to the specific systems application. The most important purpose of this testing, performed at the vendor site, is to avoid placing a defective device in a system assembly where the cost of diagnosis may far exceed the cost of incoming inspection.

2.3.5 Typical Test Flow

Actual test selection depends on the manufacturing level (processing, wafer, or package) being tested. Although some testing is done during device fabrication to

assess the integrity of the process itself, device testing is predominantly performed after the wafers have been fabricated.

- The first test, known as wafer sort or probe, isolates the potentially good devices from defective ones (Einspruch 1985, Stevens 1986). Historically, the defective dies were used to be inked using a dropper, which has been replaced by digital inking of the defective ones in a die database. After this, the wafer is scribed and cut, and the potentially good devices are packaged. The main objective of wafer sort test is to save on packaging cost by separating the good dies from the defective ones.
- After packaging, burn-in test is performed to accelerate the aging defects on packaged devices. The devices are often shaken mechanically with high g forces for a period of time. They are also subjected to high voltage and temperature stresses to accelerate the aging defects. Typically stress conditions are applied one at a time and not together. Usually, device output responses are not measured during burn-in test because the circuit may not be rated to operate under such elevated voltage or temperature conditions.
- After burn-in, the devices go through full specification testing. During class test, comprehensive testing is performed to attain high defect coverage, speed-binning through at-speed testing, and measurement of various DC and AC parameters such as I/O slew rate, standby current, PLL lock range, and lock frequency. Class test is usually quite comprehensive because often it is the last test performed by chip manufacturers before they are shipped to system manufacturers.
- At the system manufacturer end, inspection tests are conducted on the incoming devices. Manufacturers typically apply system level tests (such as high end software applications) on a sample of the incoming lot to perform a statistical study on the quality of the devices received from the fabrication house. Similar tests on a sample of parts may be applied by a chip manufacturer to ensure shipped product quality level.

Figure 2.9 summarizes different types of test methods on the basis of their objective, test metrics, type of patterns applied, and the environment variables involved as part of the testing process.

2.4 Power Delivery Issues During Test

To understand power delivery issues during various testing phases, we have to understand how the power is connected to the device under test. In Sect. 2.4.1 and 2.4.2, we thoroughly examine the power pad and packaging-related issues and power grid-related issues as manifested during testing. In Sect. 2.4.3, we discuss different sources of power supply noise (also known as *droop*) in the context of testing.

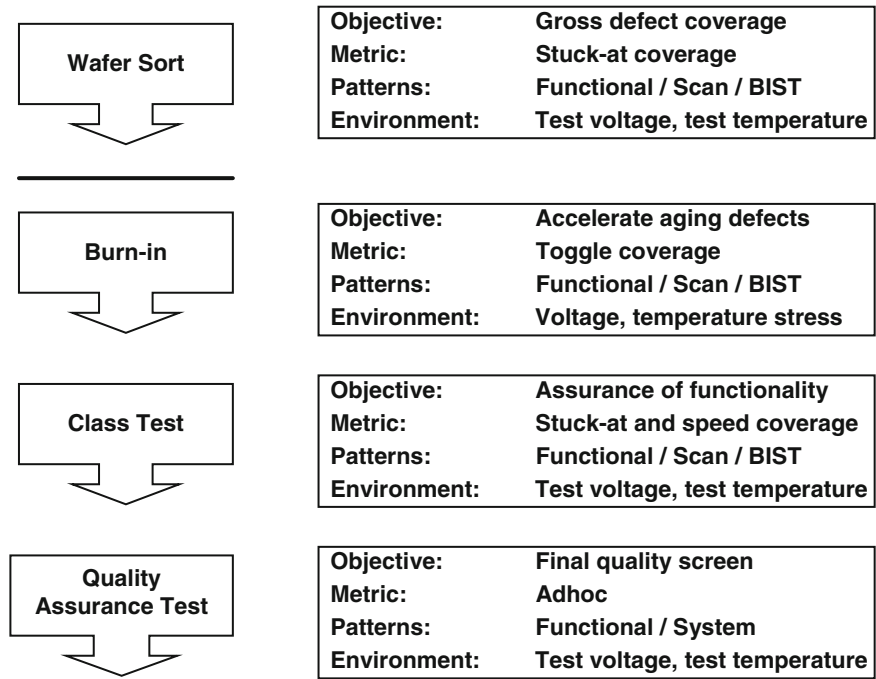


Fig. 2.9 Typical test flow

2.4.1 Packaging

Let us first investigate the role of power supply contacts between the tester and the device-under-test (DUT). To this end, we distinguish between wafer probe test and package test. To facilitate the discussion of package test, a brief description of package types is given below:

- **Wire-bonded packages:** In this technology, the pins of a bare die are situated along the perimeter and wire bonded to the package (Fig. 2.10a). Wire bonding was the only form of packaging before flip-chip technology came along.
- **Flip-chip packaging:** In this technology, the pins of a bare die are arranged as an array. The packaging substrate has a similar pin map. The die and the package are bonded together after the bare die is placed with its pin side facing down to make contact with the package substrate. This is also known as *controlled collapse chip connect* (C4) technology (Fig. 2.10b). Flip-chip technology is the dominant mode of packaging today.

DUT may not be adequately supplied with power during wafer sort test. Central problem here is that a typical C4 power contact may only be good for an average of 50 mA of current delivery to the chip, so a large array of C4 bumps is needed to supply the necessary current needed for the chip to operate at its rated power level.

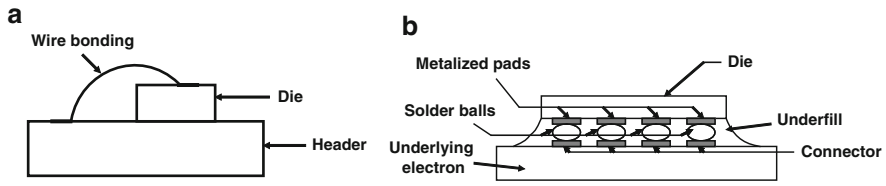


Fig. 2.10 Side-view schematic of different die mounting technology: (a) wire bonding and (b) flip-chip through C4 solder bump

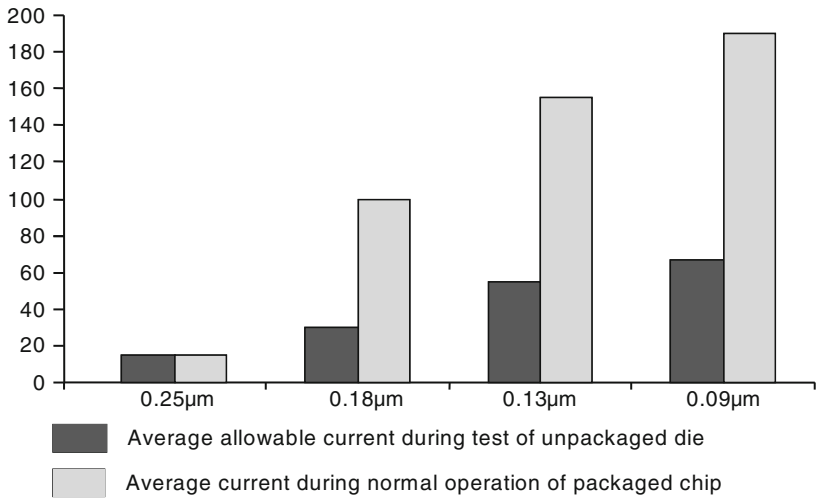


Fig. 2.11 Power availability (shown as Amps in Y-axis) during wafer testing [figure adopted from Kundu et al. (2004)]

During wafer sort test, the number of C4 pads that can be contacted by the probe pins is limited by mechanical strength of the wafer.

Since wafer thickness typically ranges from 300 to 500 μm and each probe pin applies a force of 5–10 g on the die, the number of probe contacts per unit area of the die is limited. Consequently, during wafer test all power pads may not be contacted. Power delivery constraint arising out of this limitation is shown in Fig. 2.11 (Kundu et al. 2004).

A second problem that afflicts wafer testing is the inductance of the power delivery path from the TPS to the DUT. This includes pin and C4 pad inductances, as well as inductance of wiring on the probe card and the inductance of the connectors to the tester. A large inductance on power delivery path impedes sudden changes in power consumption pattern by collapsing the supply voltage that in turn may produce false errors at the tester. The same problem can be seen in package test as well. However, package testing is usually done with chip socketing and the socket typically has large local decoupling capacitor to mitigate such problems. Similar capacitors on probe card tend to be farther away from the DUT.

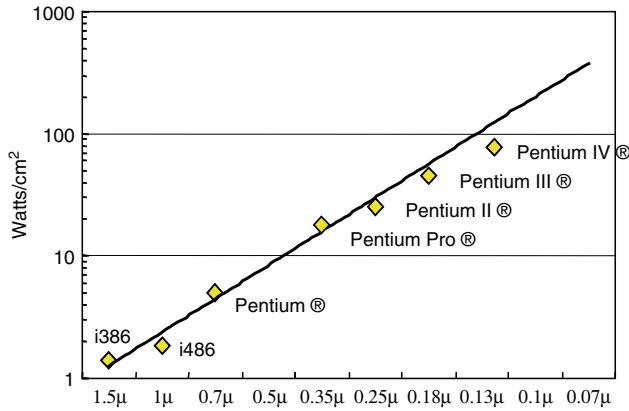


Fig. 2.12 Power density by technology [figure adopted from [Tirumurti et al. \(2004\)](#)]

2.4.2 Power Grid Issues

Increased device density due to continuous scaling of device dimensions and simultaneous performance gain has driven up the power density of high-performance computing devices such as microprocessors, graphics chips, and FPGAs. For example, in the last decade, microprocessor power density has risen by approximately 80% per technology generation, whereas power supply voltage has been scaling down by a factor of 0.8. This has led to 225% increase in current per unit area in successive generation of technologies (Fig. 2.12).

The increased current density demands greater availability of metal for power distribution. However, this demand conflicts with device density requirements. If device density increases, the device connection density will also increase, requiring more metal tracks for signal routing. Consequently, compromises are made for power delivery and power grid becomes a performance limiter. Nonuniform pattern of power consumption across a power distribution grid causes a nonuniform voltage drop. Instantaneous switching of nodes may cause localized drop in power supply voltage, which we call as *droop*. This instantaneous drop in power supply at the point of switching causes excessive delay and a path-delay problem ([Tirumurti et al. 2004](#)). There are multiple factors that contribute to power supply droop on a chip including inductance of off-chip power supply lines, inductance of package interconnects, and resistive power distribution network (PDN) on chip. The first two factors can cause large droop and must be addressed in design phase whereas the last factor has no acceptable design solution and must be addressed in test.

2.4.3 Power Supply Noise

In this subsection, we discuss the physics behind various types of power supply noise (also known as droop) in more detail. The various droop mechanisms can be

classified as low-frequency power droop, mid-frequency power droop, and high-frequency power droop. Next we describe each droop mechanism in detail.

2.4.3.1 Low-Frequency Power Droop

The current generation of microprocessors consumes 50–105 W of power (Intel White Paper 2006). At a supply voltage of 0.9–1.1 V, this translates to 45–95 amps of current. The voltage attenuation on this power line should be as small as possible. If the resistance of the power delivery line is kept in the order of $\text{m}\Omega$, the resulting IR drop will be of the order of $\text{m}\Omega \times 10^2 \text{ A} \approx 100 \text{ mV}$ or 10% of the power supply voltage. Such large drop is unacceptable. Therefore, the power delivery line needs to have even smaller resistance. Unfortunately, this tends to increase self-inductance of the power delivery line.

Let, the parasitic inductance of the interconnect be denoted by L . This inductance is associated with the power supply connector *external* to the chip as the inductance of the package pins and solders bumps. We call a sudden increase in current i demanded per unit time t (which is equivalent to a sudden increase in power consumption), a di/dt event. After a di/dt event, the DUT will see its power supply voltage V_{DD} reduced by $L di/dt$. For a current transient of 100 amp, taking place within 10^{-9} s or three cycles of a 3.3 GHz machine, this value is deleterious even for inductances L far below 1 nano-Henry (nH), whereas typical value of this inductance is 1–10 nH.

In reality, the impact of this inductance is mitigated by adding a capacitance C as shown in Fig. 2.13 to meet the short-term demand of current of the DUT during a di/dt event. The voltage droop per unit time induced by the load current is calculated as $dV/dt = i/C$ (Bakoglu 1990). The capacitor C needs to be sufficiently large to survive the $L di/dt$ effect. Typically, these transients last 50–100 ns. Even though the worst case magnitude of this drop can be severe, it is not transistor or gate specific. Usually, there is ample time to detect beginning of these events by *on-die droop detector* and respond to these events by modulating clock frequencies or flushing the pipeline and restarting computation. Thus, while these droops are severe in magnitude, they are often handled well at the design level for the functional patterns.

In test mode, self-adaptation is usually turned off because it leads to non-determinism of the output. For example, if the power supply voltage fluctuates as test patterns are being applied, due to self-governing mechanisms of a chip, accurate speed-binning cannot be performed because performance changes with supply voltage.

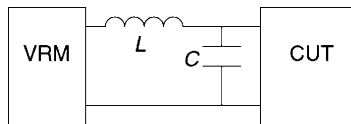


Fig. 2.13 Circuit-under-test (CUT) connected to voltage regulator module (VRM), including capacitor C and parasitic inductance of interconnect L

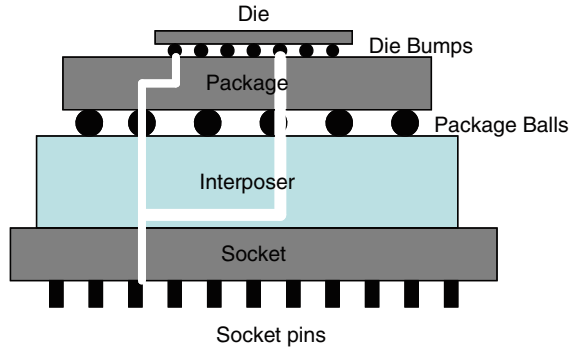


Fig. 2.14 Package showing current paths from socket pins to die bumps (figure courtesy: Intel Corporation)

On the contrary, if voltage levels are never changed, the logic associated with controlling such changes may not be tested. Thus, the onus of managing such power-level changes falls on pattern generation and test ordering mechanisms. Often such test ordering is done in an ad hoc fashion.

2.4.3.2 Mid-Frequency Power Droop

Mid-frequency voltage droop is associated with inductance at the package level. In Fig. 2.14, we show a typical package. From the socket pins to the die bumps, there are low resistance conduction paths that have reasonably high inductance (0.1–0.5 nH). During execution of instructions, if power demand shifts from one area of the die to a different area as shown with solid white line in the figure, one area of the die will experience a drop in voltage while the area where the power demand went down will experience an increase in the voltage. The package also integrates decoupling capacitance. However, owing to the scale of these interconnects, the values of both L and C are significantly smaller and the effect of voltage droop lasts 5–10 ns. For lack of a better term, droops associated with package is often called mid-frequency droop. Typically, these droops affect an entire region (integer execution unit, floating point unit, bus unit etc.) and can be addressed at the functional level by introducing multiple sensors (Clabes et al. 2004).

However, during test, if such droop is not managed well, it will lead to yield loss defined as the loss of good parts due to measurement errors during test.

2.4.3.3 High-Frequency Power Droop

High-frequency droop is associated with the PDN on the die. The PDN is usually a grid structure (Fig. 2.15). The cell library is designed with a fixed height, so that they can connect to power grid at regular points, thereby vastly simplifying the physical design process.

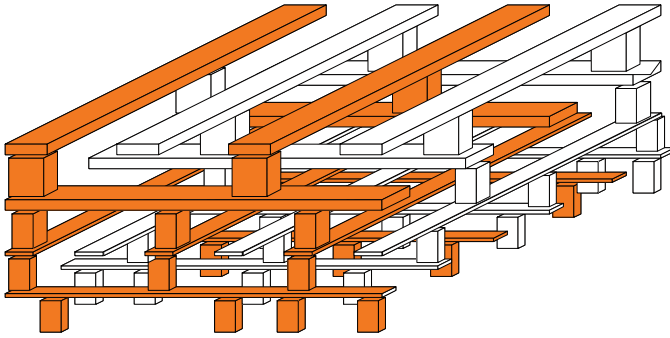


Fig. 2.15 Power distribution grid on a chip [figure adapted from Polian et al. (2006)]

The topmost metal layers (M5–M8) are often reserved for power rails and clock distribution network while lower layers are shared with logic signal lines (M2–M4). In general, the power delivery capacity of a power rail is given by its width and pitch. In microprocessor design, the width is tapered for the interconnect layers where the upper metal layers are wider. This is driven by interconnect density requirement at the lower layers and the power delivery requirement at the upper layers. There is pressure to increase the pitch of power rails as the area consumed by them is not available to logic signal lines. The vias connecting power rails of different layers transfer supply voltage from one metal layer to the next.

High-frequency power droop occurs when multiple cells drawing current from the same power grid segment suddenly increase their current demand. If the current cannot be provided quickly enough from other parts of the chip, power starvation results in a voltage drop. In contrast to low-frequency or mid-frequency power droop, this is a highly transient phenomenon lasting several hundred picoseconds.

On-die droop detector cannot be used for responding to high-frequency droops because droop detection time is usually longer than duration of the droop. Fortunately, high-frequency power droop is much smaller in magnitude. Such droops are handled in functional mode by adding a *frequency guard band*.

A similar guard band is necessary during test mode. Therefore, typically scan tests are not performed at the rated clock frequency (Xiong et al. 2008). If scan test is attempted at rated clock frequency, voltage droop due to excess switching or impedance of power supply path during test may in fact reduce the performance of a good chip below its rated level and manufacturing yield loss may occur.

2.4.3.4 Voltage Drop During At-Speed Scan

Abnormally high levels of state transitions and voltage drop during scan or BIST mode can also lead to degradation of clock frequency. It has been reported that while performing at-speed transition delay testing, fully functional devices are often discounted as “bad” causing manufacturing yield loss (Shi and Kapur 2004). During scan shift, circuit activity increases causing higher power consumption.

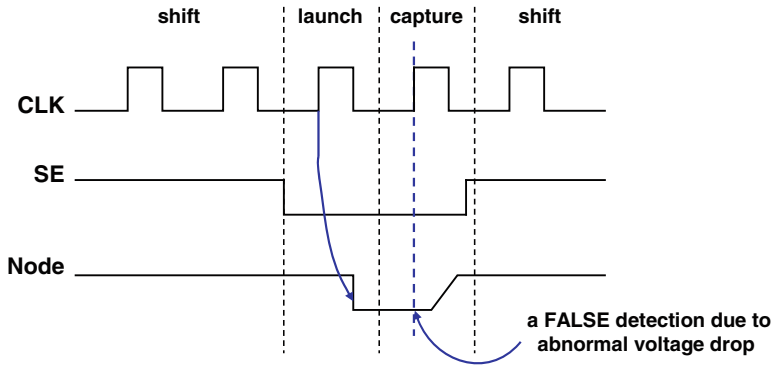


Fig. 2.16 The impact of voltage drop on shippable yield during at-speed testing [figure adopted from Shi and Kapur (2004)]

This in turn may lead to drop of power supply voltage due to IR drop where higher current or I associated with larger power dissipation causes greater voltage drop in PDN (Fig. 2.16). Such drop in voltage increases path delay requiring clock period to be stretched accordingly. If the clock period is not stretched to accommodate this increase in delay, yield loss may occur (Rearick and Rodgers 2005). IR drop not only increases path delay but also increases clock distribution latency. During structural test, a circuit toggles between system mode and scan mode. Performance of such toggle between system clock mode and scan mode may also be impacted due to reduced voltage. Thus, a chip may fail either due to excessive logic path delay or altered clock latencies or both. By contrast, in functional mode, a temporary voltage drop that increases path delay also increases clock latency that may offset increase in path delay (Wong et al. 2006). The interaction between path delay and clock latency is complex as it depends on the magnitude of each parameter as well as rise and fall times of the clock signals. However, it is safe to assume that voltage drop will increase the time it takes to toggle between scan mode and functional mode and will introduce uncertainty if the clock period itself is subject to modulation as *slow-fast-slow* as in launch off-capture or *fast-fast-slow* as in launch off-shift or any other at-speed test mechanism where the goal is to apply functional clock at speed, whereas the scan may proceed at slower speed. Such delay or uncertainty calls for capture clock to be somewhat delayed or stretched to avoid yield loss. An increase of $\sim 15\%$ in cycle time has been reported (Rearick and Rodgers 2005).

2.5 Thermal Issues During Test

The correlation between consecutive test vectors applied to a CUT is often significantly lower than that between two consecutive functional input vectors applied during its normal operation. It directly relates to higher switching activity, and therefore higher power dissipation, during test compared to normal operation mode. The

elevated levels of power dissipation during test inherently lead to higher die temperatures compared to the normal operation.

To mitigate these problems, the tests are typically applied at rates much lower than a circuit's normal clock rate in the past, since only the stuck-at fault coverage was deemed to be important. There are two recent developments in the domain of testing integrated circuits that make the power and heat dissipation during testing an extremely important issue. First, aggressive timing to improve performance of the ICs has made it essential for the tests to identify slow chips via delay testing that requires circuits to be tested at higher clock rates – if possible, at the circuit's normal clock rate (called *at-speed* testing). Second, with the advent of systems-on-chips, it is often required to test multiple cores simultaneously to reduce test application time to meet the market demand. High power and heat dissipation in neighbor cores cause undesirable thermal stress and formation of thermal hotspots. In the following subsection, we present a thorough and extensive study of various thermal hot spot-induced issues evolved during test.

Silicon die hot spots result from localized overheating, which occurs much faster than chip-wide overheating due to the nonuniform spatial on-die power distribution (Rosinger et al. 2006). Recent research supported by industrial observations suggests that spatial temperature gradients exceeding 30°C are possible even under typical operating conditions (Skadron et al. 2003), which suggest that there exist large variations in power density across the die. These gradients, especially between active and inactive blocks, are likely to increase during package testing since test power dissipation can be significantly higher compared with functional power (Pouya and Crouch 2000, Shi and Kapur 2004).

In metal oxide semiconductor (MOS) devices, there are two parameters that are predominantly sensitive to temperature: (1) the carrier's mobility μ and (2) the device threshold voltage V_t . The mobility of carriers in the channel is affected by temperature and a good approximation to model this effect is given by (Tsividis 1989):

$$\mu(T) = \mu(T_0) \left(\frac{T}{T_0} \right)^{-k_1} \quad (2.12)$$

where T is the absolute temperature of the device, T_0 is a reference absolute temperature (usually room temperature), and k_1 is a constant with values between 1.5 and 2 (Klaasen 1995). The device threshold voltage V_t exhibits a linear behavior with temperature (Klaasen and Hes 1986):

$$V_t(T) = V_t(T_0) - k_2(T - T_0) \quad (2.13)$$

where the factor k_2 is between 0.5 and 4 mV/K. The range becomes large with more heavily doped substrates and thicker oxides.

Applying these considerations to the behavior of a MOS transistor, we can predict that a temperature increment causes an increment of the drain current due to the decrease in V_t and a decrease of the drain current due to decrease in mobility. Among these two conflicting effects, the effect of mobility dominates for circuits with large overdrive voltage (which is typically the case with ultra deep submicron devices)

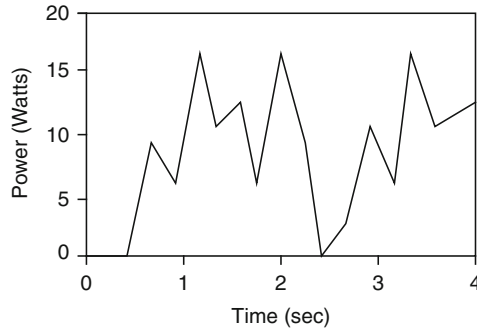


Fig. 2.17 Test patterns arranged in such a way that power cycles through high and low during the entire test period (source: Intel Technology Journal)

resulting in slowing the devices in the thermal hot spot-affected region of the chip. This will manifest as delay failures in the circuit under test causing some “good” chips to be rejected lowering the shippable yield.

In summary, local hot spots lead to (1) increased delays in gates that may register incorrectly at the tester as a defect or (2) excessive leakage that reduces electrical capacity of the local power grid that may indirectly contribute to further increased delay. Therefore, thermal hot spots during test need additional attention.

The thermal hot spot issue during test is often resolved by arranging test patterns in such a way that power is cycled high and low through the entire testing period (Fig. 2.17) so that it does not cross the temperature limits at any given time. However, applying test patterns in this way significantly increases the rate of change in supply current (di/dt). This leads to problems described in detail in Sect. 2.4.

2.6 Test Throughput Problem

Test throughput is defined as number of devices tested per test equipment over a given period. Higher the test throughput, higher is the profitability from chip manufacturing business point of view. Power consumption during testing plays a pivotal role in enhancing test throughput. In the following four subsections we discuss few of the test power related issues that directly influence test throughput.

2.6.1 Limited Power Availability During Wafer Sort Test

During wafer sort test, only a fraction of all the power pins could be used (Fig. 2.11). The fine contact pitch and the force required to create an ohmic contact with a C4 bump limits the availability of power from a mechanical point of view. Reduced power supply forces the tests to be performed at a lower frequency implying reduction in test throughput.

2.6.2 Reduction in Test Frequency During Package Test

As mentioned in Sect. 2.1, switching activity is often several times higher during testing than in normal operation mode. One ad-hoc way to reduce the dynamic power consumed during test is to lower the operating frequency, but this solution adversely affects the test throughput.

Moreover, with reduced test frequency, it takes longer to complete the entire test process. Therefore, the total energy consumed during test remains unchanged. Also, modern test requires at-speed testing to isolate slow chips, which makes testing at a lower frequency not a viable option.

2.6.3 Constraint on Simultaneous Testing of Multiple Cores

If multiple cores placed in a SoC are tested in parallel, it will reduce the overall test application time and therefore, enhance the test throughput. However, testing multiple cores in parallel may result in excessive energy dissipation and may develop thermal hot spots across the chip, which may eventually cause permanent damage to the chip. In order to control the heat dissipation during test, parallel testing of multiple sites is highly restricted contributing to a reduction in test throughput.

2.6.4 Noisy Power Supply During Wafer Sort Test

During wafer sort test, the probe card pins establish contact with the wafer metal pads, whereas the tester gets connected to the probe card connection points (Fig. 2.18). The long interconnects from the tester to the wafer metal pads offer high inductance (L). The rate of change of current (di/dt) is also high due to thermal constraint as discussed in Sect. 2.5.1; it causes *low-frequency power droop*, quantified as Ldi/dt (as discussed in detail in Sect. 2.4.3.1). To reduce this voltage noise effect, test frequency is dropped accordingly causing a drop in test throughput.

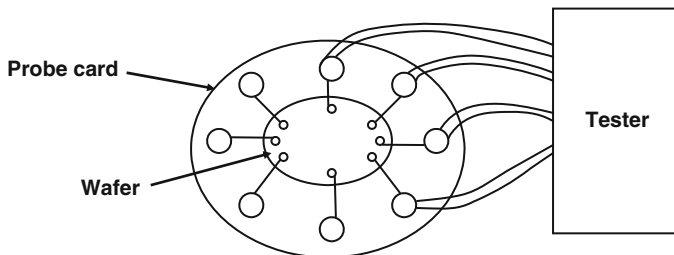


Fig. 2.18 Schematic showing connection between a wafer and the tester through a probe card

2.7 Manufacturing Yield Loss

The profitability of integrated circuits manufacturing depends heavily on the fabrication *yield*, defined as the proportion of operational circuits to the total number of fabricated circuits (Koren and Koren 1998). When a “good” circuit is falsely considered as a faulty circuit, it leads to *manufacturing yield loss*. There are several reasons behind yield loss during test, which we describe in the following five subsections.

2.7.1 ATE Timing Inaccuracy

Overall tester timing accuracy is determined by skews and parasitics between the tester and the DUT. If the test frequency is increased to the overall tester accuracy limits, it may cause significant yield loss. Unless test system timing accuracy improves in tandem with device speed, alternative test methods are necessary. For example, in PC processors the front side bus frequency has increased to 1366 MHz in recent years. When a tester is connected to front side bus, IO signals at different IOs may arrive at different times due to skew in test environment.

If the skew is greater than the rated IO period of $1/1366\text{ }\mu\text{s}$, the chip yield will go to zero. In such a scenario, IOs cannot be tested at full frequency. On the contrary, if IOs are not tested at the rated frequency, test is incomplete and alternative tests must be devised. In PC platform chips, IO wrap test or IO loopback test is often used, where signal from one IO of a chip is returned to a different IO on the same chip through a short local path allowing the chip to test its own IO speed (Fig. 2.19) (Kundu et al. 2004).

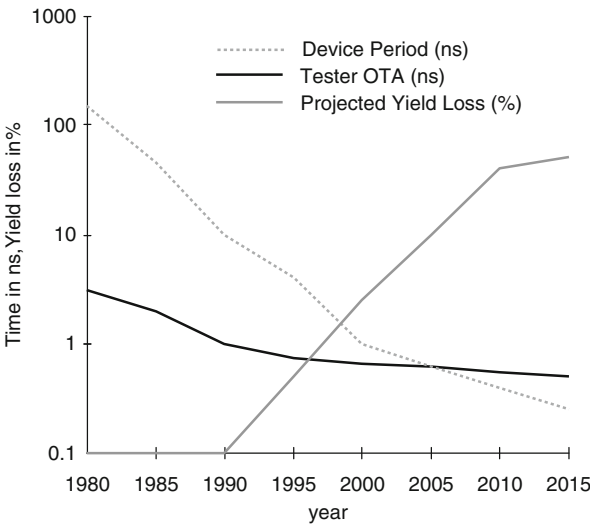


Fig. 2.19 Yield loss projection due to overall tester timing accuracy [figure adopted from Kundu et al. (2004)]

This problem is somewhat mitigated by the move to DFT-enabled IO testing. For structural testing, the only accurate timing needed is that of the system clock. Most processors or large SoCs use the tester supplied clock to generate an internal core clock, which can be many multiples of the tester clock. *On product clock generation* (OPCG) poses problems with controlling launch and capture clocks and requires additional DFT to enable such features. Circuit level verification of DFT for OPCG is also a critical issue for test. Without effective and precise launch and capture, products cannot be tested.

2.7.2 Application of Illegal Test Vectors

During structural testing, pseudorandom or deterministic patterns are applied through scan chain(s). Many of these patterns are not functional patterns and sometimes application of such a nonfunctional pattern to the DUT may perform some illegal operation from circuit perspective resulting in faulty behavior or even permanent damage of the DUT (Ogihara et al. 1983, Van der Linden et al. 1994).

The following example illustrates one such situation. Figure 2.20a shows the schematic for a 4-to-1 multiplexer (MUX), where under normal operation mode, only one input among A, B, C, or D is selected by applying appropriate selection signals (viz. S_1 , S_2 , S_3 , and S_4). If more than one selection signal becomes active, it may possibly cause a *bus contention* by driving a 0 and a 1 on the bus output at the same time (Wohl et al. 1996). Figure 2.20b shows a gate-level schematic of the 4-to-1 MUX. To make the MUX operation faster, the combinational logic for the MUX is partitioned into two parts by inserting flip-flops to store the selection signals (viz. S_1 , S_2 , S_3 , and S_4) generated by the MUX control signals (C_0 and C_1). In the second partition, these selection signals are used to select one of the four inputs as the output of the bus. We show the transistor level schematic of the second logic partition in Fig. 2.20(c). The flip-flops inserted in between the two logic partitions are part of normal scan chain during structural testing. Let us consider a bus contention caused by activating both selection signals S_1 and S_2 during scan shift operation. It will select both the inputs A and B. Now if $A = 0$ while $B = 1$, a DC path is established between V_{DD} and ground (indicated by the arrow in Fig. 2.20c). Such contending buses draw excess current that may result in a voltage drop. While a single bus contention problem may not cause a large drop in supply voltage, such contentions on a datapath consisting of wide buses may be significant. If the power supply voltage drops significantly, delays or intermediate voltage levels may cause faulty behavior in “good” chips causing manufacturing yield loss.

In summary, power supply voltage affects circuit delays as well as output voltage levels. A drop in power supply voltage increases circuit delay while its output voltage level may not saturate at expected strength levels. Alone or together, these factors contribute to manufacturing yield loss. Power delivery problems as described earlier may be artifacts of power delivery path during test, or DFT problems such as bus contention or OPCG issues or pattern related such as abrupt power level changes or contention problems.

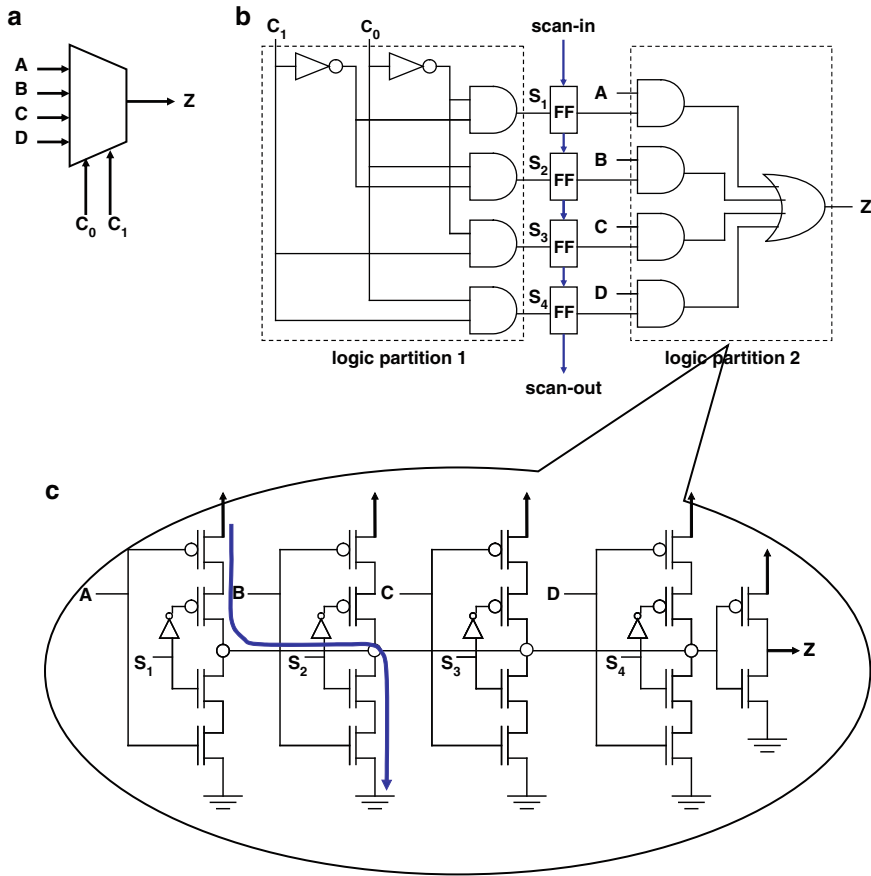


Fig. 2.20 An example illustrating a yield loss scenario due to application of an illegal test vector: (a) schematic of a 4-to-1 multiplexer, (b) gate-level schematic of the multiplexer with a scan chain partitioning the logic into two partitions, and (c) a transistor level schematic of the logic partition 2 of (b)

2.8 Test Power Metrics and Estimation

Power consumption is now considered an important constraint during test. Power estimation is required to measure the saving in power and evaluate the effectiveness of a given test power reduction technique. As both the SoC designs and the ultra deep submicron geometry becomes prevalent, larger designs, tighter timing constraints, higher operating frequencies, and lower applied voltages all affect the power consumption of silicon devices. Accurate analysis of power consumption during normal operation as well as test is necessary. Therefore, it is important to define test power metrics and their estimation.

2.8.1 Power Metrics

Following are the four major power and energy metrics that should be quantified accurately to analyze the power dissipation effects during test (Pouya and Crouch 2000):

- **Energy:** Energy is estimated as the total switching activity generated during test application. It affects the battery lifetime during power up or periodic self-test of battery-operated devices.
- **Average power:** Average power is the average distribution of power over the entire test period. Elevated average power increases the thermal load that must be vented away from the DUT to prevent structural damage (hot spots) to the silicon, bonding wires, or package.
- **Instantaneous power:** Instantaneous power is the value of power consumed at any given instant. Usually, it is defined as the power consumed right after the application of a synchronizing clock signal. Elevated instantaneous power might overload the power distribution systems of the silicon or package, causing brown-out.
- **Peak power:** The highest power value at any given instant, peak power determines the component's thermal and electrical limits and system packaging requirements. If peak power exceeds a certain limit, designers can no longer guarantee that the entire circuit will function correctly. In fact, the time window for defining peak power is related to the chip's thermal capacity, and forcing this window to one clock period is sometimes just a simplifying assumption.
- **Rate of change of power:** The highest rate of change of power affects the Ldi/dt drop and highlights deficiencies in decoupling capacitor placement or sizing. As described earlier in Sect. 2.5.3, they may cause manufacturing yield loss. Consequently, this is an important metric in characterizing power consumption during test.

2.8.2 Modeling of Power and Energy Metrics

From (2.5), we know that the average energy consumed at node i per rising transition is $C_i V_{DD}^2$, where C_i is the equivalent output capacitance and V_{DD} is the power supply voltage (Cirit 1987). Therefore, a good approximation of the energy consumed in a period is $C_i s_i V_{DD}^2$ where s_i is the number of rising transitions during the period. Nodes connected to more than one gate experience higher parasitic capacitance. On the basis of this fact, as a first approximation we assume capacitance C_i to be proportional to the fan-out count F_i of node i (Wang and Roy 1995). Therefore, an estimation of energy E_i consumed at node i during one clock period is

$$E_i = s_i F_i c_0 V_{DD}^2 \quad (2.14)$$

where c_0 is the circuit's minimum parasitic capacitance.

According to this expression, estimating energy consumption at the logic level requires the calculation of fan-out F_i and the number of rising transitions of node i , s_i over a period. Circuit topology defines the fan-out of the nodes, and a logic simulator can estimate the switchings (Girard 2002). Product $s_i F_i$ is called the weighted rising transition activity of node i and represents the only variable part in the energy consumed at node i during test application.

According to the previous formulation, the energy consumed in the circuit after application of successive input vectors $\langle V_{k-1}, V_k \rangle$ is

$$E_{vk} = c_0 V_{DD}^2 \sum_i s(i, k) F_i \quad (2.15)$$

where i ranges all the circuit's nodes and $s(i, k)$ is the number of rising transitions caused by V_k at node i .

Let us now consider a pseudorandom test sequence of length m , required to achieve the targeted fault coverage. The total energy consumed in the circuit during application of the complete test sequence is:

$$E_{\text{total}} = c_0 V_{DD}^2 \sum_{k=1}^{m-1} \sum_i s(i, k) F_i \quad (2.16)$$

By definition, the instantaneous power is the power consumed 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. Therefore, we can express the instantaneous power consumed in the circuit after application of vectors V_k as

$$P_{\text{inst}}(V_k) = \frac{E_{vk}}{t_{\text{small}}} \quad (2.17)$$

The peak power consumption corresponds to the maximum instantaneous power consumed during the test session. It, therefore, corresponds to the highest energy consumed during the same small instant of time, t_{small} . More formally we can express it as

$$P_{\text{peak}} = \max_k [P_{\text{inst}}(V_k)] = \frac{\max_k(z)}{t_{\text{small}}} \quad (2.18)$$

Finally, the average power consumed during the test session is the total energy divided by the test time:

$$P_{\text{ave}} = \frac{E_{\text{total}}}{mT} \quad (2.19)$$

where m is the number of test vectors applied during the test session.

This model for power and energy computation during test is definitely crude and simplified, but it suffices quite well for power analysis during test.

According to these expressions of power and energy consumption, and assuming a given CMOS technology and supply voltage for the circuit design, number of rising transitions s_i of a node i in the circuit is the only parameter that affects the

energy, peak power, and average power consumption. Similarly, the clock frequency used during testing affects computation of the average power. Finally, test length affects only the total energy consumption. Consequently, when deriving a solution for power and energy minimization during test, a designer or a test engineer has to keep these relationships in mind (Girard 2002).

Static power dissipation is defined as the power dissipation that occurs after all signal transitions have settled in a circuit. Therefore, static power dissipation depends only on (1) the pattern and (2) the temperature. Temperature dependence arises from transistor subthreshold leakage that increases exponentially with temperature. Consequently, static power dissipation is not a concern except in (1) I_{DDQ} testing and (2) burn-in test. In I_{DDQ} test, test application is slow and the current consumption is pattern dependent, whereas in burn-in test, static power dissipation is large due to elevated temperature. For a large circuit in nano-CMOS technology, the total leakage current does not vary greatly from pattern to pattern. Consequently, leakage current for such circuits is also defined as I_{SB} , or standby current. The main difference between I_{DDQ} and I_{SB} is that the former is pattern specific whereas the latter is not.

2.8.3 Test Power Estimation

During conventional design, power consumption in functional mode is estimated in one of the following three levels of abstraction (Najm 1994): (1) architecture-level, (2) RT-level, and/or (3) gate-level. Each one of these estimation strategies represents different tradeoffs between accuracy and estimation time (see Fig. 2.21 below).

Estimation of power consumption during test is not only required for sign-off to avoid destructive testing but also to facilitate power-aware test space exploration (during DFT or ATPG) early in the design cycle. A very inaccurate though early and fast way to estimate test power is to use architecture-level power calculators that compute switching activity factor based on architectural pattern simulation and use gate count, and various library parameters to estimate a power value (Ravi et al. 2008). However, in today's design, testing is mostly based on structural patterns

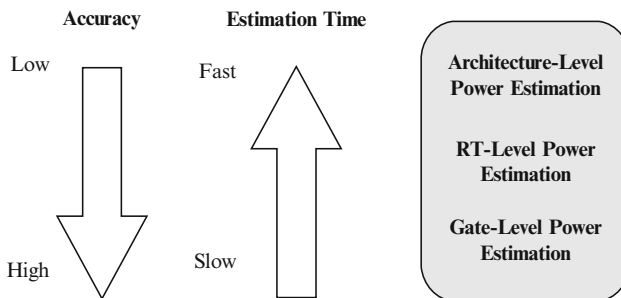


Fig. 2.21 Accuracy versus time in power estimation

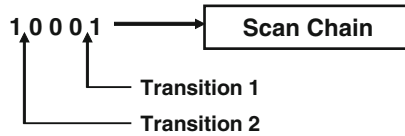


Fig. 2.22 Transitions in scan vector [figure adopted from Sankaralingam et al. (2000)]

applied through a scan chain. The architectural or RT-level designs usually do not contain any scan information that is added later in the design flow, and therefore appear only at the gate-level abstraction. Hence, gate-level test power estimator is needed. A limitation of gate-level estimation is that it is time consuming and therefore, cannot be invoked frequently early during the design cycle. Moreover, gate-level simulators are expensive in terms of memory and run time for multimillion gate SoCs. Such simulators are more suited for final analysis rather than during design iteration. RT-level test power estimators can only be used if DFT insertion and test generation can be done at the RT level (Midulla and Aktouf 2008).

Quick and approximate models of test power have also been suggested in the literature. The weighted transition metric proposed by Sankaralingam et al. (2000) is a simple and widely used model for scan testing, wherein transitions are weighted by their position in a scan pattern to provide a rough estimate of test power.

This is illustrated with an example adopted from the authors. Consider a scan vector in Fig. 2.22 consisting of two transitions. When this vector is scanned into the CUT, Transition 1 passes through the entire scan chain and toggles every flip-flop in the scan chain. On the other hand, Transition 2 toggles only the content of the first flip-flop in the scan chain, and therefore, dissipates relatively less power compared with Transition 1. In this example with five scan flip-flops, a transition in position 1 (in case of Transition 1) is considered to weigh four times more than a transition in position 4 (in case of Transition 2). The weight assigned to a transition is the difference between the size of the scan chain and the position of the transition in the scan-in vector. The total number of weighted transitions for a given scan vector can be computed as follows (Sankaralingam et al. 2000):

$$\text{Weighted transitions} = \sum (\text{Scan chain length} - \text{Transition position in vector}) \quad (2.20)$$

Although the correlation with the overall circuit test power is quite good, a drawback of this metric is that it does not provide an absolute value of test power dissipation.

2.9 Summary

Power consumption, rate of change of power consumption, and overall energy consumption are important factors during test. Availability of power during test may be limited. Abrupt changes in power consumption introduce unwanted changes to the

voltage levels in a chip. Excessive power consumption may change the operating temperatures within a chip. Such unwanted changes may invalidate tests and cause yield loss. To mitigate the impact of such changes, an array of approaches is needed. They range from modeling power delivery to heat flux analysis; test strategies, and DFT to support high-throughput power friendly tests as well as electrical verification of final test patterns to ensure that no problems are expected during testing. In this chapter, we outlined the broader set of issues and their interconnectedness. Subsequent chapters will deal with specifics.

References

- M. Abramovici, M. A. Breuer, and A. D. Friedman. "Digital Systems Testing and Testable Design". *IEEE Press*, New York City, NY, 1990
- E. Acar, R. Arunachalam, and S. R. Nassif. "Predicting Short Circuit Power from Timing Models," In *Proc IEEE Asia-South Pacific Design Automation Conference*, 277–282, 2003
- H. Bakoglu. "Circuits, Interconnections, and Packaging for VLSI". *Addison-Wesley*, Reading, MA, 1990
- M. L. Bushnell, and V. D. Agrawal. "Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits," *Kluwer Academic Publishers*, Boston, MA, 2000
- M. Cirit. "Estimating Dynamic Power Consumption of CMOS Circuits," In *Proc International Conference on Computer Aided Design (ICCAD)*, 534–537, 1987
- J. Clabes et al. "Design and Implementation of the POWER5 Microprocessor," In *Proc Design Automation Conference (DAC)*, 670–672, 2004
- V. Dabholkar, S. Chakravarty, I. Pomeranz et al. "Techniques for Minimizing Power Dissipation in Scan and Combinational Circuits during Test Application," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 17, No. 12, pp.1325–1333
- M. Drazdziulis, and P. Larsson-Edefors "A Gate Leakage Reduction Strategy for Future CMOS Circuits," In *Proc European Solid-State Circuits Conference*, pp. 317–320, 2003
- N. G. Einspruch. "VLSI Handbook," *Academic Press*, Orlando, FL, 1985
- P. Girard "Survey of Low-Power Testing of VLSI Circuits," *IEEE Design & Test of Computers*, Vol. 19, No. 3, pp. 82–92, 2002
- E. R. Hnatek. "Integrated Circuit Quality and Reliability," *Mercel Dekker*, New York City, NY, 1987
- C. Hu et al. "BSIM4 Gate Leakage Model Including Source-Drain Partition," In *Proc International Electron Device Meeting*, pp. 815–818, 2000
- A. Hirata, H. Onodera, and K. Tamaru. "Estimation of Short-Circuit Power Dissipation for Static CMOS Gates," *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences*, Vol. E79, No. A, pp. 304–311, 1996
- Intel White Paper (online resource): Intel® Multi-Core Processors: Making the Move to Quad-Core and Beyond. <http://www.intel.com/technology/architecture/downloads/quad-core-06.pdf>
- International Roadmap for Semiconductors – System Drivers (online resource): http://www.itrs.net/links/2007ITRS/2007_Chapters/2007_SystemDrivers.pdf
- F. Jensen, and N. E. Petersen. "Burn-In," *John Wiley & Sons*, Chichester, UK, 1982
- F. M. Klaasen, and W. Hes. "On the Temperature Co-efficient of MOSFET Threshold Voltage," *Solid State Electronics*, Vol. 29, no. 8, pp. 787–789, 1986
- F. M. Klaasen. "MOS Devices Modelling. In: Design of VLSI Circuits for Communications," *Prentice Hall*, Upper Saddle River, NJ, 1995
- Z. Kohavi "Switching and Finite Automata Theory," *McGraw-Hill*, New York City, NY, 1978
- I. Koren, and Z. Koren. "Defect Tolerance in VLSI Circuits: Techniques and Yield Analysis," *Proceedings of the IEEE*, vol. 86, No. 9, pp. 1819–1836, 1998

- S. Kundu, T. M. Mak, and R. Galivanche. "Trends in Manufacturing Test Methods and Their Implications," In *Proc IEEE International Test Conference*, pp. 679–687, 2004
- I. Midulla, and C. Aktouf. "Test Power Analysis at Register Transfer Level," *ASP Journal of Low Power Electronics*, Vol. 4, No. 3, pp. 402–409, 2008
- S. Mukhopadhyay, A. Raychowdhury, and K. Roy. "Accurate Estimation of Total Leakage Current in Scaled CMOS Logic Circuits Based on Compact Current Modeling," In *Proc IEEE/ACM Design Automation Conference*, pp. 169–174, 2003
- F. Najm. "A Survey of Power Estimation Techniques in VLSI Circuits," *IEEE Transactions on Very Large Scale Integrated Systems*, Vol. 2, No. 4, pp. 446–455, 1994
- N. Nicolici, and X. Wen. "Embedded Tutorial on Low Power Test," In *Proc IEEE European Test Symposium*, pp. 202–210, 2007
- T. Ogihara, S. Murai, and Y. Takamatsu et al. "Test Generation for Scan Design Circuits with Tri-state Modules and Bidirectional Terminals," In *Proc. IEEE/ACM Design Automation Conference*, pp. 71–78, 1983
- R. Pierret. "Semiconductor Device Fundamentals," Ch. 6, pp. 235–300. *Addison-Wesley*, Reading, MA, 1996
- I. Polian, A. Czutro, and S. Kundu et al. "Power Droop Testing," In *Proc. IEEE International Conference on Computer Design*, pp. 135–138, 2006
- B. Pouya, A. Crouch. "Optimization Trade-offs for Vector Volume and Test Power," In *Proc. IEEE International Test Conference*, pp. 873–881, 2000
- J. M. Rabaey, A. Chandrakasan, and B. Nikolic. "Digital Integrated Circuits: A Design Perspective," *Prentice Hall*, Upper Saddle River, NJ, 1996
- S. Ravi, S. Parekhji, and J. Saxena. "Low Power Test for Nanometer System-on-Chips (SoCs)," *ASP Journal of Low Power Electronics*, Vol. 4, No. 1, pp. 81–100, 2008
- J. Rearick, and R. Rodgers. "Calibrating Clock Stretch During AC Scan Testing," In *Proc. International test Conference*, 2005
- P. Rosinger, B. M. Al-Hashimi, and K. Chakrabarty. "Thermal-Safe Test Scheduling for Core-Based System-on-Chip Integrated Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 25, No. 11, pp. 2502–2512, 2006
- K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand. "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, Vol. 91, No. 2, pp. 305–327, 2003
- R. Sankaralingam, R. Oruganti, and N. A. Toub. "Static Compaction Techniques to Control Scan Vector Power Dissipation," In *Proc. IEEE VLSI Test Symposium*, pp. 35–42, 2000
- C. Shi, and R. Kapur. "How power aware test improves reliability and yield," *EETimes*. <http://www.eetimes.com/news/design/features/showArticle.jhtml?articleId = 47208594&kc = 4235>. Accessed 21 November 2008
- K. Skadron, M. Stan, and W. Huang et al. "Temperature-Aware Microarchitecture. In *Proc. International Symposium on Computer Architecture*, pp. 2–13, 2003
- A. K. Stevens. "Introduction to Component Testing," *Addison-Wesley*, Reading, MA, 1986
- Y. Taur, and T. H. Ning. "Fundamentals of Modern VLSI Devices," *Cambridge University Press*, New York City, NY, 1998
- C. Tirumurti, S. Kundu, S. Sur-Kolay et al. "A Modeling Approach for Addressing Power Supply Switching Noise Related Failures of Integrated Circuits," In *Proc. IEEE Design, Automation, and Test in Europe Conference*, pp. 1078–1083, 2004
- Y. P. Tsividis. "Operation and modeling of the MOS Transistor," *McGraw-Hill*, New York City, NY, 1989
- J. T. H. Van der Linden, M. H. Konijnenburg, and A. J. Van de Goor. "Test Generation and Three-State Elements, Busses and Bidirectionals," In *Proc IEEE VLSI Test Symposium*, pp. 114–121, 1994
- H. J. M. Veendrick. "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid State Circuits*, Vol. 19, No. 4, pp. 468–473, 1984

- S. Vemuri, and N. Scheinberg. "Short-Circuit Power Dissipation Estimation for CMOS Logic Gates," *IEEE Transactions on Circuits and Systems-I*, Vol. 41, No. 11, pp. 762–765, 1994
- C. Y. Wang, and K. Roy. "Maximum Power Estimation for CMOS Circuits Using Deterministic and Statistical Approaches," In *Proc. IEEE VLSI Conference*, pp. 364–369, 1995
- S. Wang, and S. K. Gupta. "ATPG for Heat Dissipation Minimization During Test Application," *IEEE Transactions on Computers*, Vol. 47, No. 2, pp. 256–262, 1998
- N. H. E. Weste, and K. Eshraghian. "Principles of CMOS VLSI Design: A Systems Perspective," *Addison-Wesley*, Reading, MA, 1988
- P. Wohl, J. Waicukauski, and M. Graf. "Testing "Untestable" Faults in Three-State Circuits," In *Proc. VLSI Test Symposium*, pp. 324–331, 1996
- K. L. Wong, T. R.-Arabi, and M. Ma et al. "Enhancing Microprocessor Immunity to Power Supply Noise with Clock-Data Compensation," *IEEE Journal of Solid State Circuits*, Vol. 41, No. 4, pp. 749–758, 2006
- K. N. Yang, H. T. Huang, and M. J. Chen et al. "Characterization and Modeling of Edge Direct Tunneling (EDT) Leakage in Ultra Thin Gate Oxide MOSFETs," *IEEE Transactions on Electron Devices*, Vol. 48, No. 6, pp. 1159–1164, 2001
- J. Xiong, V. Zolotov, and C. Visweswariah et al. "Optimal Margin Computation for At-Speed Test," In *Proc. IEEE Design, Automation and Test in Europe Conference*, pp. 622–627, 2008

Power-Aware Testing and Test Strategies for Low Power
Devices

Girard, P.; Nicolici, N.; Wen, X. (Eds.)

2010, XXI, 363 p., Hardcover

ISBN: 978-1-4419-0927-5