Heat and Power Management for High Performance Integrated Circuits

by

Arman Vassighi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2004

I hereby declare that I am the sole author of this thesis.
I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Arman Vassighi

I further authorize the University of Waterloo to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Arman Vassighi

Abstract

Among all the issues that CMOS scaling has faced, increased power consumption in general and leakage power in particular are among the most important issues that the VLSI designers have to address. Due to the strong correlation between power consumption and operating temperature, increased power consumption compromises the reliability, functionality and performance of the circuits, either during chip normal operating condition or during test and reliability screening. Thermal modeling of high-performance circuits and systems is a crucial factor in order to achieve reliable and power-saving designs. The VLSI community currently lacks a way to model temperature at any level of design other than low-level circuits. The accuracy of thermal modeling has a substantial effect on the accuracy of thermal management studies of the processor architecture. Without this essential modeling capability, architecture researchers are limited to inaccurate estimation techniques, which will not be suitable for the thermal management of high performance circuits. In this thesis some of these issues are discussed and new models and associated CAD tools are developed. Various techniques at the circuit and system levels are explored.

In this thesis, a technique for junction temperature estimation is developed. Using this technique, the increase in the normalized junction temperature with scaling under nominal and burn-in conditions was predicted. This thesis also provide a new insight into the concept of thermal runaway and how it may best be avoided. Finally an electro-thermal tool was developed to study the low temperature operation of the high performance processors, while incorporating different techniques at circuit, and system levels. In this tool all the physical parameters of the chip at device, circuit and system level was incorporated and the tool was calibrated to an actual microprocessor.

Acknowledgements

I am truly honored to have been working under the supervision of Dr. Manoj Sachdev. His continued encouragement and guidance, in addition to his moral support and understanding are very much appreciated.

I am indebted to my Ph.D. committee members Dr. J. Hamel, Dr. A. Hassan, Dr. K. Ponnambalam, and my external examiner Dr. B. Cockburn for their helpful comments and constructive critiques.

I will also like to thank Dr. A. Keshavarzi from Intel Corporation for his support and guidance during my Ph.D. program.

No words can be used to thank God for his blessings, my father, my dearest mother, my brothers, and my wife for their continues love and support in all my endeavors. Especially my mother whom her endless encouragement and support, made this journey possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Evolution of CMOS Technology

The ability to improve performance with reduced power consumption per logic gate made CMOS the dominant technology for integrated circuits. Transistor scaling is the primary factor driving speed performance improvement in both microprocessors and memories. Historically, CMOS technology scaling per technology node has:

- Reduced the gate delay by 30% allowing an increase in maximum clock frequency of 43%.
- Doubled the device density.
- Reduced the parasitic capacitance by 30%.
- Reduced the energy and active power per transition by 65% and 50%, respectively [15][72][81].

To achieve this, transistor width, length, and oxide dimensions were scaled down by 30%. As a result, the chip area was decreased by 50% for the same number of transistors, and total parasitic capacitance was decreased by 30%. Recent data on microprocessor operating frequencies show this trend [72]. Figure 1.1 shows the evolution of Intel microprocessor operating clock frequency and gate delays per clock since 1987.

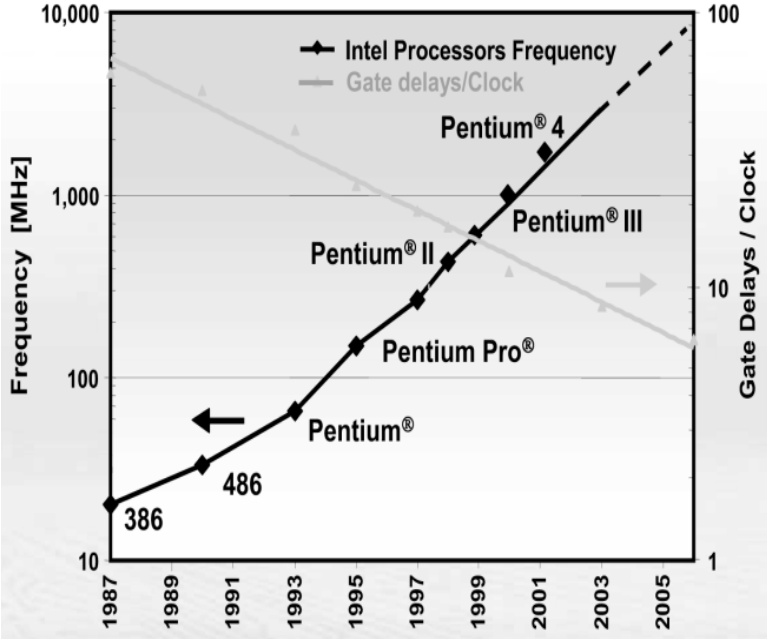


Figure 1.1: Processor frequency trend adopted from [72].

The classic scaling described above has not been strictly followed in commercial products. Classic scaling has served as an essential blueprint describing the major features observed over the period from roughly 1981 to 2001. Figure 1.2 shows a collection of published industry results for electrically-equivalent transistor gate-oxide thickness (T_{OX}), threshold voltage (V_T), and power supply voltage (V_{DD}), all plotted against the reported gate length (L_{GATE}). Dashed lines show the classic scaling trajectories for these parameters as well. Taking gate length as a measure of the lithography scale, one can immediately see that V_{DD}, V_T, and, to a lesser extent, T_{OX} have decreased more slowly than L_{GATE}, while I_{DSAT} has actually increased rather than remaining fixed (as in classic scaling). The right-hand side of the figure shows the same V_T and T_{OX} data as the left-hand side, except with V_{DD} as the abscissa. Note that T_{OX} and V_T fall relatively close to scaling in proportion to V_{DD} (as they would in classic scaling). This suggests that the deviations from classic scaling have been driven primarily by V_{DD}, which has itself decreased more slowly than L_{GATE}. In the early part of this time span ($1\mu m$ to $0.5\mu m$), a reluctance to leave the widely accepted industry-standard V_{DD}=5.0V, inherited from Transistor-Transistor Logic (TTL), substantially retarded V_{DD} reduction. As the transition to a $3.3V$ standard gained momentum, an increased emphasis on performance and power resulted in circuit board designs with more flexibility for V_{DD}; these, in turn, allowed CMOS process technology developers the freedom to optimize V_{DD} scaling for power and performance to a greater degree.

A given technology point defined by specific values of T_{OX} and L_{GATE} will nearly always deliver greater performance as V_{DD} is increased (roughly in direct proportion to V_{DD}), so as gate dielectric learning in the industry accelerated, the acceptable ratio of V_{DD}/T_{OX} increased steadily in this next era, giving rise to a continued mismatch in L_{GATE} and T_{OX} reduction rates. Thus, V_{DD} continued to decrease more slowly than L_{GATE}.

The other item of note in Figure 1.2 is the behavior of V_T. A large scatter in V_T is seen, due in part to variability in reporting practices (nominal vs. fast process, V_T definition, etc.). However, to a good approximation, V_T scaled in proportion to V_{DD}. This is probably largely a consequence of practical CMOS device and circuit considerations, including circuit stability, noise immunity, and the engineering of short channel effects to acceptable levels of control. These observed behaviors are seen to give rise to a number of practical problems that pose challenges to further CMOS scaling [62].

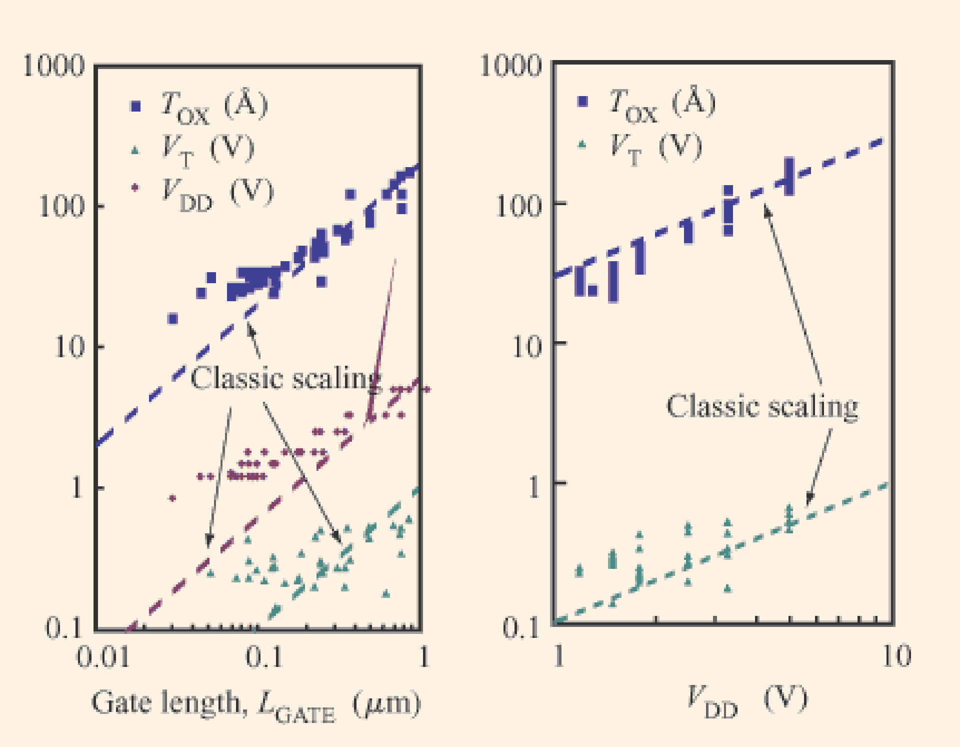


Figure 1.2: Published industry trends (data points) are compared to classic scaling (dashed curves) [62].

The supply voltage and transistor threshold voltages (V_T) are also reduced by 30% under the constant electric field scaling scenario. V_T must be scaled to maintain a sufficient gate overdrive $(V_{DD} - V_T)^n$ where n varies between 1 and 2 [3]. V_T scaling has serious impact on increased leakage current. Sub-threshold leakage is an inverse exponential function of V_T, so that the chip leakage power increases exponentially with technology scaling.

1.2 Issues in Product Quality and Reliability

The reliability and quality of electronic products counts for 50% of the total cost of the product. High reliability products not only ensure that the field maintenance cost is as low as possible, but also ensure the functionality of the systems that are intolerable to failures. The quality of the products is examined by extensive test procedures prior to the shipment of the product. Reliability is the probability that an object will function satisfactorily under given conditions for a given time without failure. Reliability engineering is becoming increasingly important for the competitive success of industry. Effective reliability estimation and improvement requires a fairly sophisticated set of skills. The scaling down of devices in advanced VLSI circuits has created major reliability problems. Some of the mechanisms that are affecting the reliability of electronic devices are latchup, electrostatic discharge (ESD), hot carrier effects, thin dielectric breakdown, and electromigration. Screening and accelerated tests are carried to detect early life failures and estimate the mean time to failure of the product. Optimization of these tests is an important factor in maximizing the yield while maintaining the effectiveness of the tests.

1.3 Thermal Issues in High Performance Processors

Power dissipation limits have become a major constraint in the design and thermal management of high performance circuits such as processors. Off state leakage current is an increasing percentage of the total current at the 130-nm and sub-100 nm nodes under nominal conditions and is expected to increase further with scaling. For 130-nm technology the leakage power is 20% to 50% of the total power in high performance microprocessors, such as Intel's Pentium 4, and is expected to increase to even more than 50% for sub-100 nm technologies. Figure 1.3 illustrates the increase in total power consumption and the increasing percentage of off state leakage current at the 130 nm and 100 nm nodes.

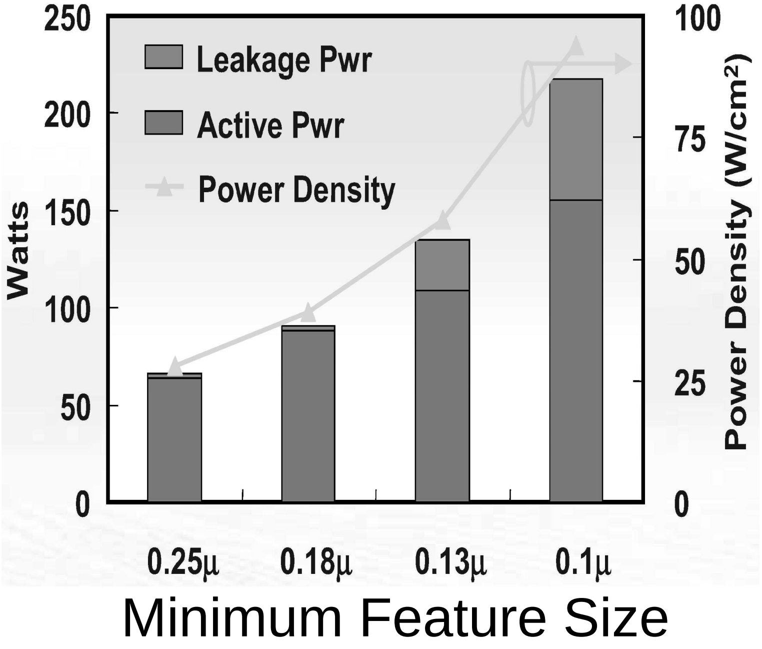


Figure 1.3: Power density trend adopted from [72]. Assumptions: 15 mm^2 die, 1.5x frequency increase per generation.

Moreover, in a reliability screening environment (e.g. burn-in) where ICs are tested under voltage and temperature stress, the ratio of leakage to active power becomes adverse and increases the probability of thermal runaway. These issues must be addressed at the architectural, circuit design and packaging levels. In other words, thermal management in high performance VLSI circuits will become an integral part of design, test, and manufacturing.

1.4 Thesis Overview

This thesis is organized in six chapters. After this introduction, chapter 2 describes burn-in as a reliability screening test and discusses the burn-in issues with respect to technology scaling. In chapter 3, after reviewing the concept of thermal resistance in CMOS, a novel technique is introduced to estimate the junction temperature in normal and burn-in conditions. Later in this chapter burn-in optimization with respect to reliability and yield is discussed. Thermal runaway as a threat to the yield of VLSI chips during burn-in is discussed in detail in chapter 4. A self-consistent electro-thermal modeling tool was developed to study the tradeoffs of low temperature operation. Chapter 5 describes this model and presents the result of a low temperature operation study. Finally, the conclusion and future work are presented in chapter 6.

Chapter 2

CMOS IC Technology Scaling and Its Impact on Burn-in

The total power consumption of high performance microprocessors increases with scaling. Off state leakage current is an increasing percentage of the total current at the 130 nm and sub-100 nm nodes under nominal conditions. The ratio of leakage to active power becomes adverse under burn-in conditions and the off state leakage can become the dominant power. Typically, clock frequencies are kept in the tens of MHz range during burn-in resulting in a substantial reduction in active power. On the other hand, the voltage and temperature stresses cause the off state leakage to be the dominant power component.

Stressing during burn-in accelerates the defect mechanisms responsible for early life failures. Thermal and voltage stresses increase the junction temperature resulting in accelerated aging. Elevated junction temperature, in turn, causes leakage to further increase. In many situations, this may result in positive feedback leading to thermal runaway. Such situations are more likely to occur as technology is scaled down to the nano meter regime. Thermal runaway increases the cost of burn-in dramatically. Other than thermal runaway, another issue with over stressing the chip is that the useful life of the chip will be shorter than it was planned for and this raises the long-term reliability issues. Hence, the temperature and voltage stress must be carefully optimized and tailored for any chip exposed to burn-in conditions.

2.1 What is Burn-in?

Component failure mechanisms and failure phenomena have been studied for a long time. Through experience and much data gathered by researchers and practitioners, component failure rates have been shown to follow the traditional bathtub curve.

The traditional bathtub curve (Figure 2.1) depicts component life in three stages. During the first stage, the failure rate begins high and decreases rapidly with time. This stage is known as the infant mortality period, and it has a decreasing failure rate (DFR). The infant mortality is mostly due to latent reliability defects. The infant mortality period is followed by a steady-state failure rate period, which is usually long and has a constant failure rate (CFR). This second stage is called the normal operating life and this is the period that the device will operate under normal conditions. Finally, the curve ends in the third stage, a period of wearout with an increasing failure rate (IFR). This is the period of aging. It is common for electronic devices to follow the traditional bathtub failure pattern.

2.1.1 Infant Mortality

Generally, a reasonable definition of the infant mortality period includes all failures prior to the normal operating period of the device life with its relatively stable and low steady state failure rate. The infant mortality period of the life cycle results from failures in a weak sub-population of the devices. The percentage of the weak sub-population (usually

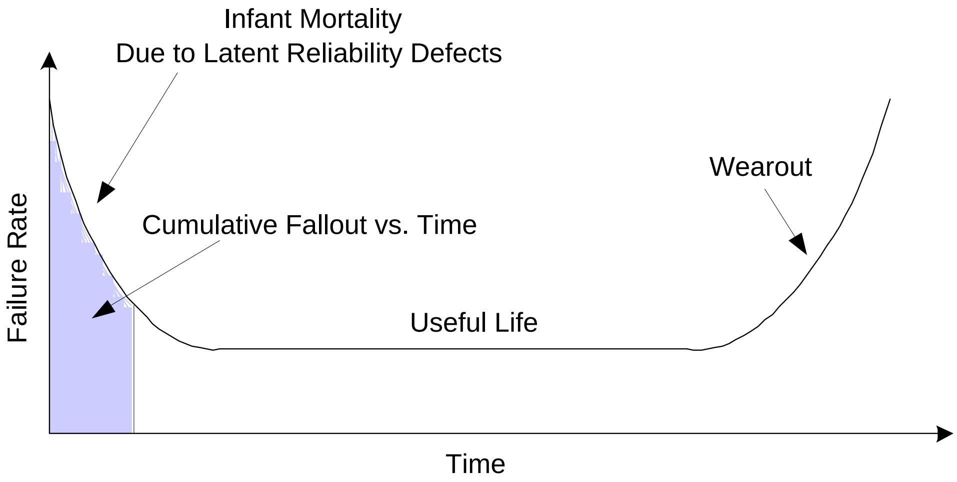


Figure 2.1: Bathtub curve [53].

a small percentage), varies with the component type and the manufacturing lot, even for the same manufacturer. Factors contributing to the infant mortality include:

- Surface anomalies, for example: corrosion, contamination, and electromigration.
- Moisture entry.
- Quality defectives such as poor workmanship, irregularities, and process deviations.
- Electrostatic discharge.
- Random failures.

The above problems cannot be entirely eliminated, although good design and manufacturing help considerably. The distribution function of the infant mortality stage has been modeled as a Weibull distribution, a log-normal distribution, a non-homogeneous Poisson process, and a empirical distribution.

2.1.2 Why Burn-in?

In principle, burn-in is a process of eliminating defective parts from the production batch. The final tests that separate functional ICs from nonfunctional ones, in effect, are screening tests. However, ICs with defects that function marginally may not be eliminated by such screens and would end up in the field and begin to fail very early in the life of the system. The failure of these weak parts gives rise to the high initial failure rates commonly observed in the infant mortality period. A process of detection and elimination of such devices is called reliability screening.

Burn-in is a reliability screening method which requires acceleration of the mechanisms that give rise to infant mortality. The concept of the screening process is to accelerate the failures until the surviving population would begin its operational life with the low failure rates corresponding to the middle steady region of the bathtub curve. Temperature, voltage bias, and a combination of the two are often used as stresses to accelerate failures. The test conditions are selected depending on the nature and degree of the failure mechanisms causing infant mortality.

2.1.3 Burn-in Procedures

Traditionally, the burn-in procedure is executed prior to a final functional test procedure that weeds out the parts that have impaired functionality and/or high leakage current from the stresses during burn-in. Burn-in systems are designed to test hundreds of units in parallel over a period of many hours with operating frequencies in the tens of MHz range. There are three basic implementation methods for burn-in:

- Final package burn-in, where dies are packaged into their final destination packages and are subjected to burn-in at temperatures within the package thermal design constraints.

- Die level burn-in, where dies are placed into temporary carriers before they are actually packaged into their final form, thus reducing the cost of waste associated with added packaging.
- Wafer level burn-in (WLBI), where dies are tested while still in wafer form.

The last method potentially offers the greatest cost savings by eliminating the packaging waste cost. The first method offers the most reliable final product since package-related reliability issues are also taken into account. However, this method is expensive since fewer packaged devices can be burnt-in simultaneously, and post burn-in loss includes packaging cost. WLBI is relatively inexpensive, but it results in a relatively less reliable product since packaging related reliability issues are not addressed. Finally, the die-level burn-in with temporary carriers offers a compromise between the other two methods.

2.1.4 Static and Dynamic Burn-in

In static burn-in, dies are loaded into burn-in board (BIB) sockets; the BIBs are placed in the burn-in oven. The burn-in system applies power to the devices and heats them to $125°C - 150°C$ for periods ranging from 12 to 24 hours. In static burn-in, the device under test (DUT) is powered but inputs are not toggled.

Dynamic burn-in mimics the static burn-in process, but also stimulates the DUT address, data, and clock inputs at a reduced rate (10-30 MHz) determined by the relatively cheap electronics of the burn-in tester. Under dynamic conditions, circuit nodes are toggled ensuring that voltage stress is applied to various transistors. Neither static nor dynamic burn-in monitors the DUT responses during the stress. Weak die destroyed by the burn-in process are not detected until a subsequent functional test stage. ”Intelligent” burn-in systems not only apply power and signals to DUTs; they also monitor the DUT outputs.

The Test During Burn-in (TDBI) method can guarantee that devices undergoing burn-in are indeed powered and that input test vectors are being applied. In addition, TDBI can perform some test functions. Detailed information about different burn-in methods and features of burn-in ovens can be found elsewhere [76][19][39].

2.2 Reliability Issues and Acceleration Factors

The effects of temperature and V_{DD} on microelectronic devices are often assessed by accelerated tests carried out at high temperature and voltage to generate reliability failures in a reasonable time period. Burn-in is often used as a reliability screen to weed out infant mortalities. Weak gate oxides are one of the major components of such failures. These failures are accelerated due to elevated electric field and temperature. Several dielectric breakdown models exist in the literature that can describe intrinsic as well as the defect-related breakdown. In the next subsections, we consider some widely used models. It is apparent that electric field and junction temperature influence the time to breakdown of a gate oxide. Metal failures are another typical reliability failure mechanism activated by burn-in. Most metal failures are due to electromigration [63][35] or stress voiding [35].

2.2.1 Time-Dependent Dielectric Breakdown Models (TDDB) - Gate Oxide Breakdown Models

The fundamental physical mechanisms of gate oxide breakdown are divided into two groups: intrinsic and extrinsic oxide breakdown mechanisms. The intrinsic oxide breakdown and wearout refers to defect-free oxide. The failure mechanism can be defined at the critical density of accumulated charge traps in the gate oxide through which a conductive path is formed from one interface to the other. The extrinsic breakdown refers to defects in the oxide whose failure mechanisms are the result of plasma damage, mechanical stress inside of oxide film, contamination, hot carrier damage, or oxide damage by ion implantation. The extrinsic damages in gate oxide typically appear during relatively short time burn-in testing (e.g. 12 hours). Both breakdown mechanisms appear during burn-in as well as life in testing [56][92].

The E and 1/E models are widely used in intrinsic gate oxide reliability predictions for oxide thickness greater than $50A^{\circ}$. Both models have a known physical basis. The E-model is expressed as:

$$t_{bd} = A.exp(-\gamma E).exp(\frac{E_a}{kT_j}) \tag{2.1}$$

where t_{bd} is the time to breakdown, A is a constant for a given technology, γ is the field acceleration parameter, E is the oxide field, E_a is the thermal activation energy, k is Boltzman's constant, and T_j is the junction temperature (K). Based on the E-model, increasing electric field across the gate oxide will decrease the time to break down.

On the other hand, the researchers have argued that the breakdown process is a current driven process, thus t_{bd} should be dependent on $1/E$. The $1/E$ model predicts:

$$t_{bd} = \tau_{\circ}.exp(\frac{G}{E}).exp(\frac{E_a}{kT_j}) \tag{2.2}$$

where $\tau_{\circ}$ and G are constants, E is the oxide electric field, E_a is the activation energy, and T_j is the junction temperature. The 1/E model implies that the dielectric will not degrade in the absence of electric field. The 1/E model ignores important thermal/diffusional processes that are known to degrade all materials over time, even in the absence of an electric field.

To increase the drive current and to control the short channel effects, the gate oxide thickness should decrease at each technology node. Experimental measurements of the time to breakdown of ultra thin gate oxides with thickness less than $40A^{\circ}$ show that the conventional E and 1/E TDDB models cannot provide the necessary accuracy for calculation and prediction [75]. Hence, starting from about 130-nm CMOS technology (T_{OX} range is about 26-31 A°) a new TDDB model has been proposed [75][60]. Experiments show that the generation rate of stress-induced leakage current (SILC) and charge to breakdown (q_{bd}) in ultra thin oxides is controlled by the gate voltage rather than the electric field. This model (Equation 2.3) includes the gate oxide thickness (T_{OX}) and the gate voltage (V_G) [57].

$$t_{bd} = \tau_{\circ}.exp[\gamma(\alpha.T_{ox} + \frac{E_a}{kT_j} - V_G)] \tag{2.3}$$

where γ is the acceleration factor, E_a is the activation energy, α is the oxide thickness acceleration factor, T_{ox} is a constant for a given technology, and T_j is the average junction temperature [57].

2.2.2 Electromigration (EM)

Interconnect EM is the movement of metal atoms in the direction of electron flow due to momentum transfer from electrons to the metal ions under thermal and voltage stresses. EM is usually modeled by the empirical Black's formula [12], that relates the Mean-Time-To-Failure (MTTF) to the stressing conditions and is given as:

$$MTTF = A.J^{-n}.exp(\frac{E_a}{kT_j}) \tag{2.4}$$

where A is a process constant dependent on the material and geometry of the metal strip, n is a current exponential factor, T_j is the absolute junction (chip) temperature, k is Boltzmann's constant, E_a is the activation energy and J is the current density. The activation energy for Al-Cu metal is in the range of 0.76-0.86 eV [50] and the activation energy for Cu interconnections can vary widely from 0.7-0.9 eV to 1.0 eV. The lifetime of interconnect decreases with reductions in the line width [37]. The accuracy of lifetime prediction is strongly dependent on the accuracy of the junction temperature measurement during the acceleration testing.

2.2.3 Temperature and Voltage Acceleration Factor Models

Several industrial reliability standards are based on temperature and voltage acceleration factor models. The MIL-HDBK-217F US military standard defines the temperature acceleration factor as [48]:

$$\pi_T = 0.1.exp(-A(\frac{1}{T_j} - \frac{1}{298})) \tag{2.5}$$

where A is a constant and T_j is the junction temperature (K). Similarly, the voltage acceleration factor is defined in the CNET reliability procedure as [17]:

$$\pi_V = A_3.exp(A_4.V_A.(\frac{T_j}{298})) \tag{2.6}$$

where A_3 and A_4 are constants, V_A is the applied voltage, and T_j is the junction temperature (K).

These reliability-prediction models show that the average junction (chip) temperature is a fundamental parameter, and should be accurately estimated for each technology generation. To do this, we must understand the properties of new materials and processes used for implementing VLSIs.

2.3 Technology Scaling and Burn-in

Traditionally burn-in is used to accelerate the early life of an IC to detect the infant mortalities. Figure 2.2 shows the bathtub curve for three different technologies. As we scale to smaller channel lengths, the useful life period of the chip shrinks from more than 7 years (10 years for technologies above $0.25\mu m$) in $0.18\mu m$ technology to less than 7 years in $0.10\mu m$ technology [53]. This is due to the increasing junction temperature as we scale to deep sub-micron technologies. The increase in junction temperature arises from higher operating frequencies and consequently higher dynamic power and also increased static power which is due to elevation in leakage power. The useful life period of the IC is shrinking due to higher junction temperature operation, higher hot electron injection due to higher I_{D-sat}, and consequently more probable gate oxide wear-out. On the other hand, electromigration at higher junction temperatures and higher current densities will cause interconnect failures at higher rates. Therefore it is important to carefully optimize the burn-in conditions to avoid over stressing the ICs in scaled technologies. Over stressing the chip in the burn-in environment will further reduce its useful life period and increase post burn-in fallouts. As shown in Figure 2.2, the failure rate in the early stage (0-30 days) of the life of an IC is 500 DPM (devices per million) and within the first year it is 200 FIT, where 1 FIT is 1 failure per 10^9 hours, or approximately 1 failure per 100,000 years (114,155 years to be precise). The failure rate during the useful operational life of the IC is constant but during the aging stage starts to increase due to intrinsic defects (electromigration, hot electron injection, etc) with a failure rate of less than 0.1%.

2.3.1 Active Power

The power of an integrated circuit (IC), for a fixed operating voltage and temperature, increases linearly with the clock frequency f (the frequency of a master signal with which all operations must be synchronized) driving the IC. Extrapolation of the power vs. frequency response down to a frequency of zero (which may be realized in a sleep mode) yields a non-zero power, which is referred to as the static power, P_{static}. The component of power which

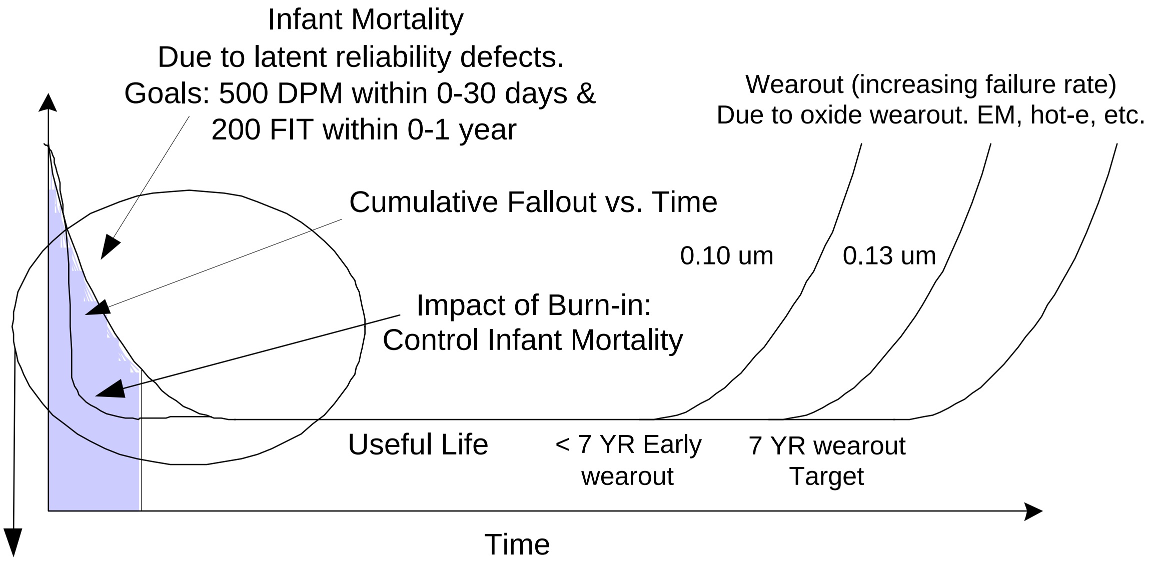

Infant Mortality
Due to latent reliability defects.
Goals: 500 DPM within 0-30 days &
200 FIT within 0-1 year
Wearout (increasing failure rate)
Due to oxide wearout. EM, hot-e, etc.
Cumulative Fallout vs. Time
0.10 um
0.13 um
Impact of Burn-in:
Control Infant Mortality
Useful Life
< 7 YR Early
wearout
7 YR wearout
Target
Time

Figure 2.2: Bathtub curve shrinks with technology scaling due to higher junction temperature [53].

is proportional to the frequency is referred to as the dynamic power, $P_{dynamic}$. The dynamic power is due primarily to the charging and discharging of capacitances in the IC, and can be represented by an effective switching capacitance, C, via the well known relationship,

$$P_{dynamic} = C.V_{DD}^2.f \tag{2.7}$$

In this equation C does not necessarily represent the actual total capacitance being switched by the chip since many of the circuits may be switching at some fraction of f (or, for that matter, at some multiple of f). Furthermore, another source of active power, sometimes referred to as short-circuit, shoot-through, or crossover power, is also lumped into C. This short-circuit power is due to current which completes a path from the power supply node to ground directly through a network of n-type and p-type FETs during the short but finite time interval when the gates are close to $V_{DD}/2$, and hence both n- and p-type FETs are in a conducting state. Typically this component represents several percent of the active power.

2.3.2 Static Power and Scaling

The standby current density increases exponentially as the length scale is decreased. This follows from the demand that V_T decrease with V_{DD}, together with the observation that

$$I_{off} \sim I_o.exp(-V_T \times \frac{q_e}{nkT}) \tag{2.8}$$

where q_e is the electronic charge, k is Boltzmann's constant, and T is the absolute temperature. This I_{off} dependence is simply a thermodynamic relationship describing the minority-carrier population (the inversion channel) as a function of temperature and the energy level in the silicon. While $n \sim 1.4$ for practical designs today, the theoretical lower bound for any FET, even decreasing n to 1, provides only minor reductions to I_{off}, given the low values of V_T ($\sim 0.4V$) at present. Furthermore, in the most recent generations of CMOS, the rate of tunneling of electrons and holes through gate oxides has increased to a point at which these currents must also be considered. These currents cause an additional power demand in the operation of CMOS which is often referred to as static power, since, unlike switching, or active power, static power is dissipated by all CMOS circuits all of the time, whether or not they are actively switching.

Figure 2.3 illustrates the static power trend based on subthreshold currents calculated from the industry trends of V_T, all for a junction temperature of $T_j = 25°C$. More practical values of T_j only serve to exacerbate this situation, with the off current of MOSFETs rising nearly two times for each $10°C$ increase in T_j. For reference, the active power density is shown in Figure 2.3 in the same scale to illustrate that the subthreshold component of power dissipation is emerging to compete with the long battled active power component for even the most power-tolerant, high speed CMOS applications. Empirical extrapolation (dashed curves) suggest that sub-threshold power will equal active power at $L_{gate} = 20nm$ and this point will be encountered closer to $L_{gate} = 50nm$ at elevated temperatures [62].

2.3.3 Static Power under Stress Conditions

As we scale the transistor down to the deep sub-micron regime, its off state leakage increases significantly. A linear reduction in transistor threshold voltage with technology scaling results in an exponential increase in its leakage. This leakage is further increased under voltage and temperature stress conditions. The total leakage of a transistor in $0.18\mu m$ technology as a function of temperature and voltage stress is shown in Fig. 2.4.

The leakage power doubles for every $10°C$ increase in junction temperature. Since the burn-in test is performed at a reduced frequency (tens of Megahertz), the dynamic power reduces from 75%-80% of total power to a negligible amount when compared to static

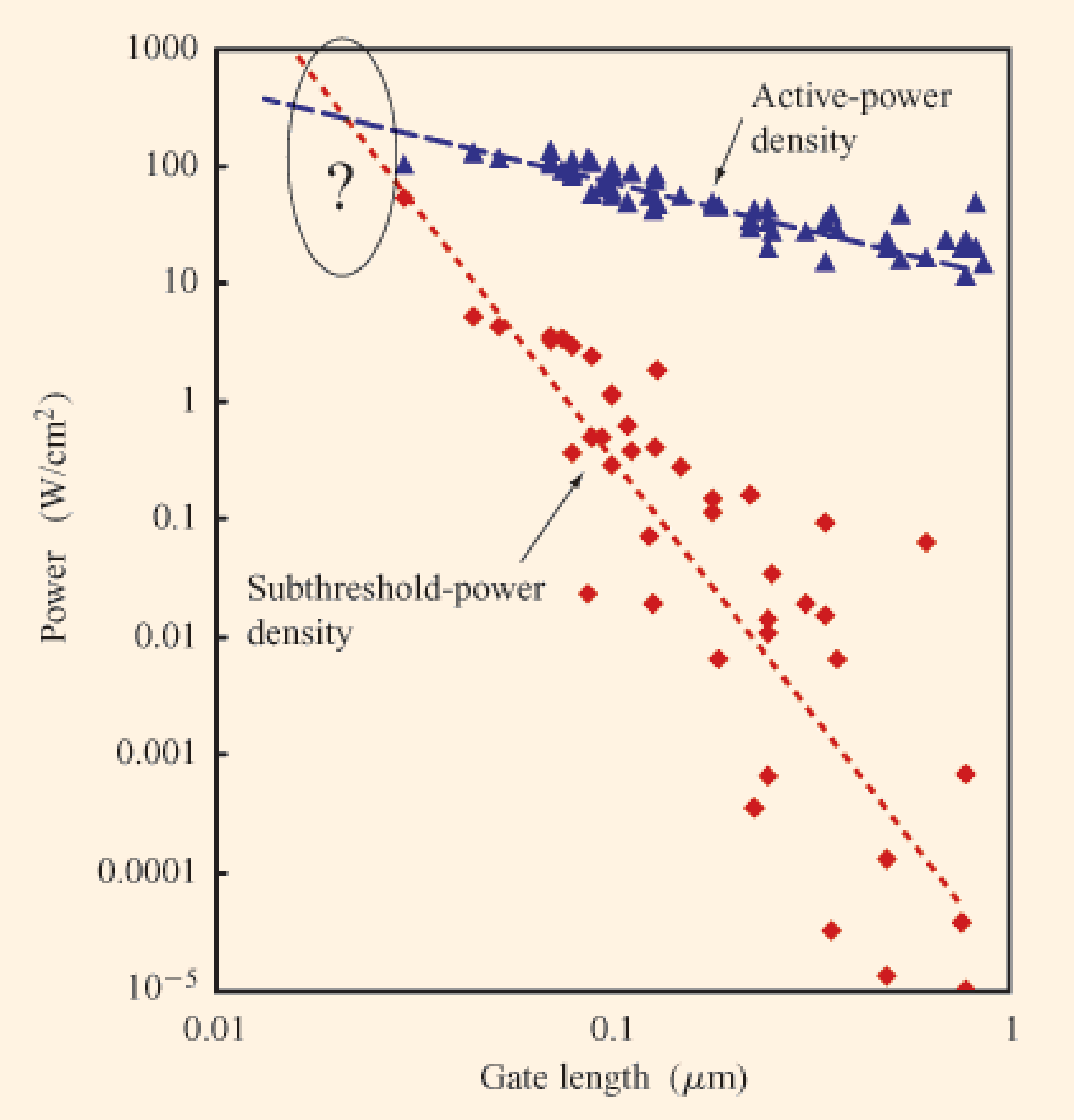


Figure 2.3: Active power density and sub-threshold leakage power density trends, calculated from industry trends, are plotted vs. L_{gate} (points) for a junction temperature of $25°C$ [62].

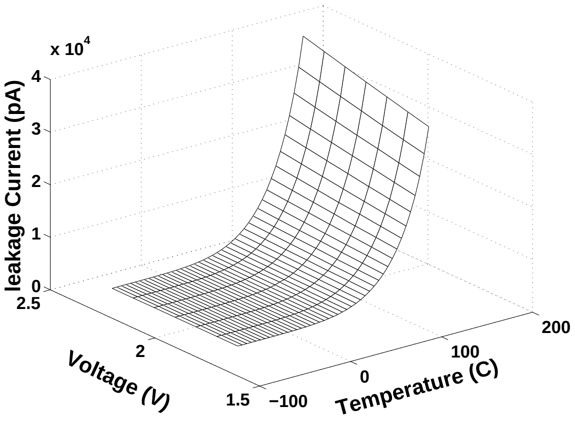


Figure 2.4: SPICE simulation of transistor leakage as a function of voltage and temperature in TSMC 0.18-μm technology.

power. Figure 2.5 shows that under stress conditions, different leakage components which account for 20% to 25% of the total power under nominal conditions in 0.13μm technology, are increased due to temperature and voltage stress and account for almost all the power under stress conditions. It must be noted that some of these leakage components are mainly sensitive to voltage stress, like gate leakage, and some of them are temperature and voltage sensitive, like subthreshold leakage.

2.4 Low Power Circuit Techniques

Much research has been carried out on low power and leakage current reduction [71]. The power consumption in CMOS circuits can be divided into dynamic and static categories. Despite increasing leakage currents with scaling, the dynamic power constitutes the majority of power consumption under normal operational conditions. However, under burn-in conditions, the leakage power becomes significantly large while the operational frequency

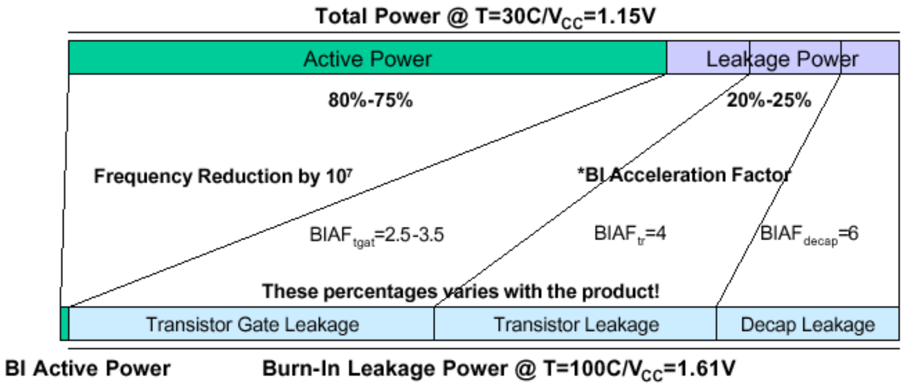


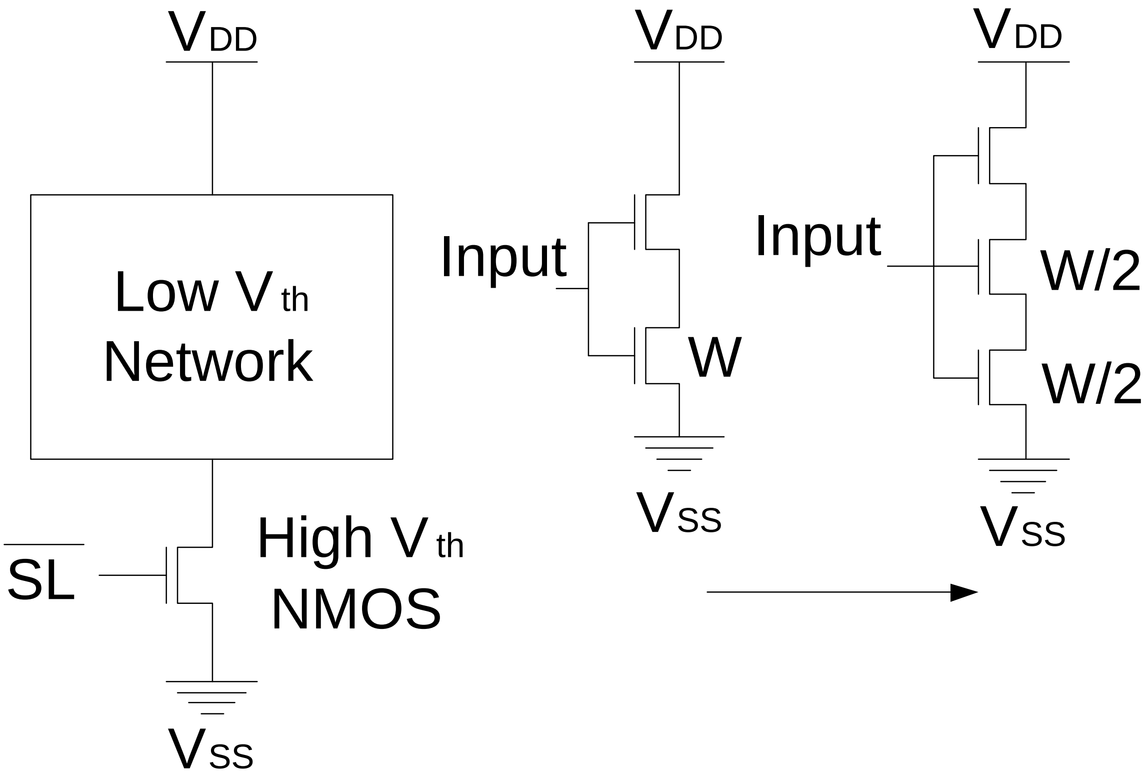
Figure 2.5: Leakage power in burn-in conditions dominates the total power of the chip [53].

is reduced drastically. Consequently, the static power component is the dominant part of the total power consumption.

Several circuit techniques have been used to reduce the background leakage current [16]. Some of these techniques can be used during burn-in to restrict the increase in leakage current and are described below.

2.4.1 Multi-threshold Logic

This technique adjusts the high performance critical path transistors by using low V_T while non-critical paths are implemented with high V_T transistors. Hence, performance and power objectives are achieved at the cost of additional process complexity. Wei et al., reported a reduction of more than 80% in leakage power while meeting the performance objectives by using a dual V_T technology [88]. Alternatively, a high V_T transistor can be placed between the power supply/ground and the high performance circuit or block (Figure 2.6(a)). In the active mode, the high V_T transistors are on and since their on-resistance is low, the performance impact is minimal. In the standby mode, the high V_T transistor is



(a) MTCMOS

(b) Two Stack under iso input

Figure 2.6: (a) MTCMOS and (b) Stack effect.

off, and hence the leakage is limited to the leakage of a high V_T transistor [89].

Traditionally, multi-threshold transistors are realized through different doses of threshold adjust ion implantations. Adjusting the threshold voltages can also be done by depositing two different oxide thicknesses or by different channel lengths [88].

2.4.2 Stack Effect

Another solution to the increasing leakage places a non-stacked transistor on a stack of two transistors without affecting the input load [58]. It has been shown that stacking two off transistors significantly reduces the sub-threshold leakage compared to a single off transistor (Figure 2.6(b))). The drawback of this technique is the increased delay. This delay increase is comparable to high V_T logic implementation in a dual V_T technology.

A significantly large fraction of the non-critical path implemented with this technique shows minimal performance degradation while reducing the sub-threshold leakage. The stack forcing technique can be either used in conjunction with dual V_T or with a single V_T technology [58].

2.4.3 Reverse Body Bias (RBB)

This is another technique to reduce leakage current during active operation, burn-in, as well as in standby mode. During active operation, RBB is applied to the idle portion of the chip to reduce the overall chip leakage power without impacting the performance. Since the chip operational frequency is very low during burn-in, RBB can be applied to the whole chip simultaneously.

Although, increasing RBB reduces the weak inversion current monotonically, the junction leakage component increases with larger RBB due to the GIDL effect. An optimal point is achieved where any further increase in RBB does not produce an overall subthreshold current reduction. The effectiveness of RBB is diminishing with scaling. Keshvarzi et al. showed that the maximum leakage reduction through RBB is from 4-5X in 180 nm technology to 2-3X in 130 nm technology [44].

2.4.4 Conditional Keepers

Degradation of dynamic circuit functionality is a problem during burn-in testing because of high leakage in stress conditions. To overcome this problem, a keeper technique was proposed that is active during burn-in, and is inactive during normal operation. Consequently, the dynamic circuit remains functional under burn-in without relaxing the maximum burn-in stress and without any significant performance degradation under normal operating conditions [5].

The elevated temperature and voltage exponentially increases the leakage current. The large leakage current can discharge dynamic nodes resulting in incorrect operation of dynamic circuits. Conditional burn-in keepers are designed to ensure the functionality of sub-130 nm dynamic circuits. The conditional keeper technique uses an extra keeper for the burn-in mode to compensate for the higher leakage during burn-in. Figure 2.7 shows this technique. Transistor M1 is the standard keeper, while transistor M2 is the burn-in keeper. M2 is off in the normal operation and turns on for the burn-in mode using a burn-in signal through the NAND gate.

2.5 Burn-in Elimination

The elimination of burn-in by an alternate screening method has been a long sought after goal. However, despite the expense, mechanical and EOS/ESD damage to the burn-in

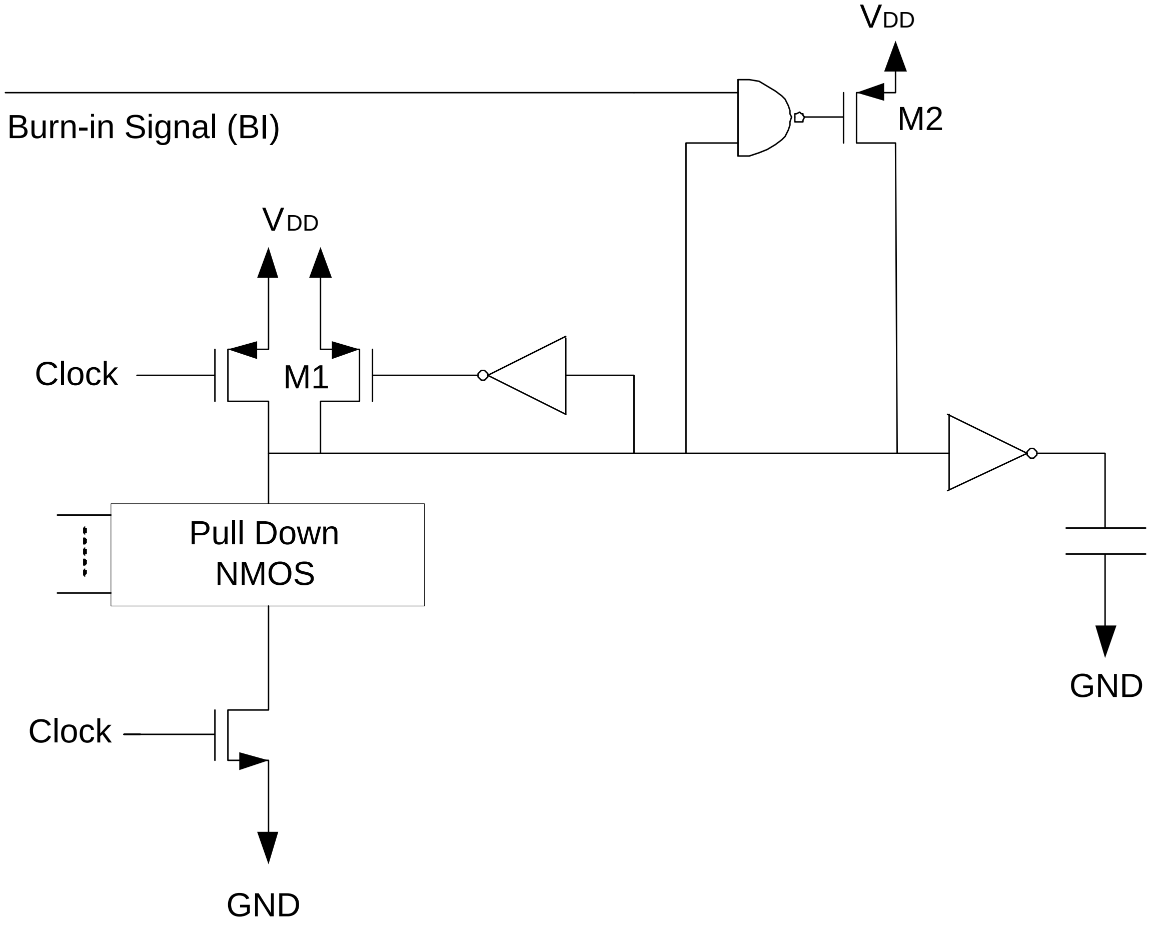


Figure 2.7: Burn-in conditional keeper in dynamic circuits [5].

parts, and lengthened time to market, burnt-in parts typically achieve a better quality measure than non-burnt-in parts. The negative features of burn-in stimulated a search for screening methods that might achieve the same lowering of DPM levels of shipped parts. In the pre-nanometer technologies, where transistor channel lengths were above $0.35\mu m$, the I_{DDQ} test was reported by several companies as successful at eliminating or reducing burn-in [36][43][11][87][55]. Intel reported experiments on several thousand ICs and found that I_{DDQ}, when combined with a short high voltage stress on the parts, yielded near zero DPM outgoing quality levels [36]. Kawasaki Steel reported a similar study using several hundreds of thousands of parts showing that I_{DDQ} screens could eliminate burn-in [43]. LSI Logic and Philips Semiconductors reported similar success with I_{DDQ} screening to eliminate burn-in [11][87]. McEuen of Ford Microelectronics reported that nominal voltage I_{DDQ} testing enabled reduction of burn-in failures by 51% [55].

However, one caveat of these reports was that I_{DDQ} screening was successful in burn-in elimination only if the manufacturing quality levels were high. I_{DDQ} could not eliminate burn-in on rogue lots. This obstacle was overcome in a study funded jointly by Sandia National Labs and the Sematech organization [68]. The experiment used 3,495 parts in a dynamic burn-in that separated the parts into a control sample, a 7 V stress sample, and an 8 V stress sample. 40,000 I_{DDQ} measurements were taken per die during the control and voltage stress sample tests. I_{DDQ} test limits were set tightly at the $+/-3\sigma$ levels from the mean plus a tester noise guard-band. Figure 2.8 summarizes the prediction of functional failure during burn-in from pre-burn-in I_{DDQ} test data. The I_{DDQ} screen predicted that I_{DDQ} testing would detect 50% of the control parts (5 V), 54% of the 7 V stressed parts, and 77% of the 8 V stressed parts. DPM of the data showed that the DPM level of the control group was 1.75 times larger than the 8 V stressed sample. Cost models also showed economic justification of the I_{DDQ} test in eliminating burn-in. A test methods study was

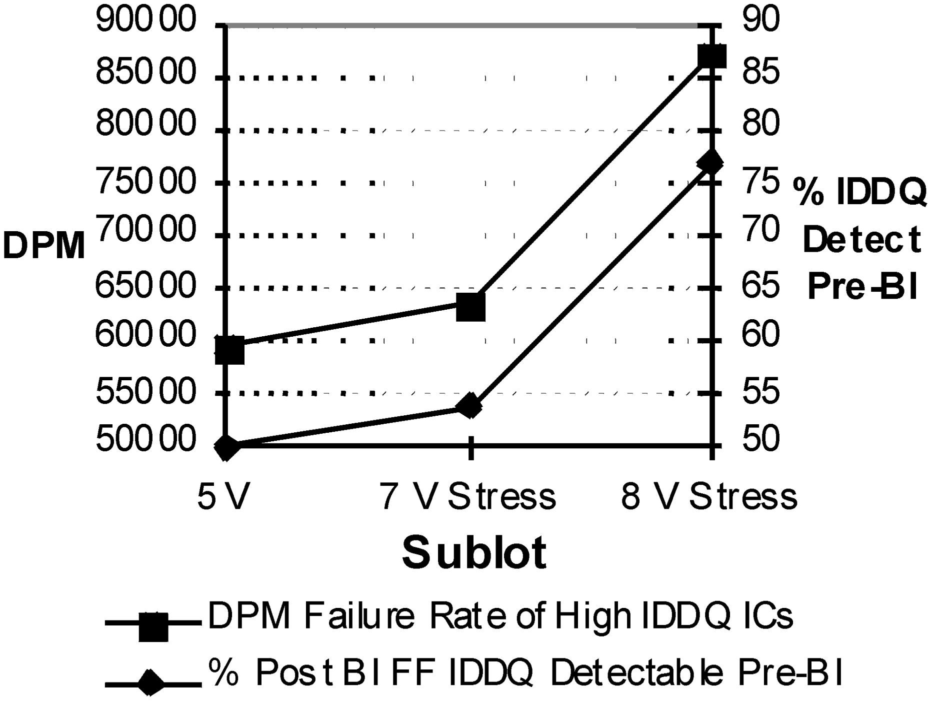


Figure 2.8: I_{DDQ} detection of burn-in functional failures and the defect level of ICs that failed only I_{DDQ} tests [68].

also funded by Sematech with IBM. This is the only study to date that stated that I_{DDQ} testing can not replace burn-in [61]. However, no explanation was given as to why the data contradicted the several reports that it would, and no burn-in data were given.

While these experiments demonstrated that parametric measurements could be used to eliminate burn-in, they were done on long channel transistor ICs whose background noise levels obscured sensitive I_{DDQ} or other parametric measurements. An important question now is how does I_{DDQ} or other parametric measurements perform for nanometer CMOS ICs. There are two public reports of success. The first was at a burn-in panel at the International Reliability Physics Symposium (IRPS) in 2001 [3]. Panelists from five major companies said that if the manufacturing quality of the lots could be measured as high, then parametric screens could achieve BI elimination. They stressed that this approach did not work if the quality levels were not high.

The second report on nano-technology parts came from a team from LSI Logic and Portland State University [21][51][74]. They reported parametric screening of outlier parts using post-test statistical processing methods on the whole wafer data. The technique measures statistics of neighboring or other die locations on the wafer to determine I_{DDQ} and V_{DDMin} (lowest functional voltage V_{DD}) test limits. These studied reported the application of post-test statistics to burn-in elimination, but did not specifically report the burn-in elimination data. The severe problems that nanometer ICs present to burn-in make these parametric screening techniques of great interest.

2.6 Conclusions

Burn-in is a quality improvement procedure widely used for high performance and high volume products. This chapter provides an overview of CMOS technology scaling and its impact on burn-in.

Smaller geometries, increased transistor leakage, and larger integration are resulting in higher junction temperatures and self-heating. Elevated junction temperature, in turn, causes leakage to increase further. The effects of higher temperature on leakage and the possibility of over stressing the chip during burn-in was discussed. Also different circuit techniques to reduce leakage power were reviewed.

Significant research has been carried out towards burn-in elimination. For long channel devices, several companies have reported burn-in elimination with I_{DDQ} under controlled process conditions. However, it appears to be difficult to eliminate burn-in for deep submicron technologies.

Chapter 3

Junction Temperature Projections for Deep Sub-micron Technologies

Several techniques can estimate junction temperature. One method directly measures junction temperature with thermal sensors at several on-chip locations during normal and burn-in conditions [31][33]. Another method uses chip level 3D electro-thermal simulators that can find the steady-state CMOS VLSI chip temperature profile at the corresponding circuit performance [18][80]. However, thermal sensors are relatively large devices, and accurate prediction requires a number of them placed on the IC. Sensors require calibration. Gerosa, et. al., reported a $0.2mm^2$ thermal sensor with a sensing range of $0 - 128°C$ and a 5-bit resolution ($4°C$) [67]. Thermal sensors can only be used for verification, and one may have to use other techniques for prediction and estimation. 3D electo-thermal simulators cannot be used for large-scale integrated circuits such as microprocessors because of long simulation time. The simulation time of a 2D Discrete Cosine Transformation (DCT) chip (107,832 transistors, 8 MHz) was reported at 12 hours [80].

In this chapter, a method for average junction temperature (T_j) estimation that can be used for normal and burn-in operating conditions is proposed. The method can predict the impact of technology scaling on junction temperature. The packaging issues, such as the thermal impedance of the package and other such factors were not considered. In this work the focus was on the intrinsic die behavior under the burn-in and normal conditions since package thermal properties tend to be user-specific.

3.1 Semiconductor Thermal Resistance Models

The Arrhenius model predicts that the failure rate of integrated circuits is an inverse exponential function of the junction temperature. A small increase of $10-15°C$ in junction temperature may result in $\sim 2X$ reduction in the life span of the device [86]. While T_a represents the ambient temperature for an IC, the relationship between ambient and average junction temperature for a VLSI is often described as in [79]:

$$T_j = T_a + P_{chip} \times R_{ja} \tag{3.1}$$

where T_a is the ambient temperature, P_{chip} is the total power dissipation of the chip, and R_{ja} is the junction-to-ambient thermal resistance. The impact of technology scaling on Equation 3.1 must be analyzed to estimate the average junction temperature for several technologies. In this work the power dissipation and thermal resistance change with technology scaling were investigated in order to predict how these parameters will change.

The initial investigations on technology scaling and thermal resistance were carried out on bipolar transistors. For these devices, the thermal resistance was estimated as in [40]:

$$R_{ja} = \frac{1}{4K(L \times W)^{0.5}} \tag{3.2}$$

where K is the thermal conductivity of silicon, $(L \times W)$ is the emitter size, and R_{ja} is the thermal resistance ($°C/mW$). It was shown that the thermal resistance increased as the emitter size was reduced. Recently, a relationship between the thermal resistance of a MOSFET and its geometrical parameters was derived using a 3-D heat flow equation [69].

$$R_{ja} = \frac{1}{2\pi K}[\frac{1}{L}ln(\frac{L+(W^2+L^2)^{0.5}}{-L+(W^2+L^2)^{0.5}}) + \frac{1}{W}ln(\frac{W+(W^2+L^2)^{0.5}}{-W+(W^2+L^2)^{0.5}})] \qquad (3.3)$$

where K is the thermal conductivity of silicon ($K = 1.5 \times 10 - 4W/mm^{\circ}C$ [70]), W and L are channel geometry parameters. The thermal conductivity of silicon has a temperature dependence described in [13].

The temperature dependence of silicon thermal conductivity is more important in silicon on insulator (SOI) technologies where self-heating contributes to a rise in junction temperature. So, our calculations assumed that the thermal resistance of silicon was temperature independent [69][70]. Equation 3.3 was used for the thermal resistance calculations for MOSFETs in different CMOS technologies.

3.2 Effect of Scaling on Junction Temperature in Normal and Burn-in Conditions

In low-power applications, the power supply voltage and transistor sizing are scaled more aggressively to minimize the power consumption [22][14]. The transistor threshold voltage in low power ICs is typically higher than for high-performance ICs to suppress the sub-threshold leakage. At the same time, the speed relative to the high-performance case should not degrade more by than 1.5X [22]. Our focus will be on high performance applications where dynamic and static power consumption are relatively high and pose a serious reliability threat.

3.2.1 Estimation of Junction Temperature Increase with Technology Scaling at Normal Conditions

F_{max} is defined as the maximum toggle frequency of an inverter in a given technology. The dynamic power consumption calculation under normal operating conditions was done at 70% of F_{max}. HSPICE simulations were carried out with BSIM model level 49. Transistor models for a $0.13\mu m$ CMOS technology were taken from United Microelectronics Corporation (UMC). Transistor models for other CMOS technologies were adapted from the Taiwan Semiconductor Manufacturing Corporation (TSMC). The simulation results and transistor sizes are given in Table 3.1. The inverter's load was the standard load element (n-MOSFET) used by TSMC for inverter ring-oscillator simulations. The load element sizes were taken from the TSMC and UMC SPICE models file specified for each of analyzed CMOS technologies. The International Technology Road map for Semiconductors (ITRS) 2002 [26] indicates that scaling down of device sizes is still in progress. Planar type transistors with 15-30 nm gate lengths have already been demonstrated [25]. However, 90-100 nm CMOS technology is currently the state-of-the-art for production of

CMOS Tech./V_{DD}	N-MOSFET	P-MOSFET	N-MOSFET	F_{max}	$F_{operating}$
	W/L	W/L	load W/L		$= 0.7F_{max}$
$(\mu m/V)$	$(\mu m/\mu m)$	$(\mu m/\mu m)$	$(\mu m/\mu m)$	(MHz)	(MHz)
0.35/3.3	4.0/0.35	10.0/0.35	3.0/3.5	1450	1015
0.25/2.5	2.86/0.25	7.14/0.25	2.15/2.5	1950	1365
0.18/1.8	2.06/0.18	5.14/0.18	1.55/1.8	2300	1610
0.13/1.2	1.49/0.13	3.71/0.13	1.12/1.3	4000	2800

Table 3.1: Simulated CMOS inverter parameters and F_{max}.

microprocessors and SRAM chips [38][64][27]. Therefore, we included the 90 nm CMOS technology node in our study of burn-in testing. The effective channel length of transistors for this technology was assumed to be 55-65 nm.

The total power consumption of an inverter toggling at $0.7F_{max}$ in four different technologies is simulated, with results given in Table 3.1. The thermal resistance of an average transistor was computed from Equation 3.3. The average size of a transistor was estimated by averaging the NMOS and PMOS transistor widths. As the transistor dimensions are reduced, the thermal resistance increases. Figure 3.1 illustrates inverter power dissipation at an operating frequency of $0.7F_{max}$ and the thermal resistance of an average transistor as functions of technology. Owing to lack of access to 90 nm CMOS technology, an alternative method was utilized to obtain the inverter power and thermal resistance estimates in Figure 3.1. For the 1.0 V, 90 nm CMOS technology, the ITRS predicts the transistor density in a microprocessor chip to be about 0.27 millions/mm^2. It is assumed that the transistor density is doubled with technology scaling for each new process generation. An industrial estimate of the power density of a microprocessor chip, implemented in 90 nm technology, is approximately $0.5W/mm^2$ [15][64][27]. Power density is defined as the power dissipated by the chip per unit area under nominal frequency and normal operating conditions. Using these assumptions we can estimate the inverter power dissipation at normal operating conditions ($V_{DD} = 1\ V$, $T = 25°C$) and speed (Figure 3.1).

The scaling scenario of transistor sizes in a CMOS inverter was extended to 90 nm CMOS technology to calculate the thermal resistance. Transistor sizes of P-MOSFET (W/L)=3.0/0.1 and N-MOSFET (W/L)=1.0/0.1 were used. The calculated transistor thermal resistance for 90 nm technology using Equation 3.3 is shown in Figure 3.1.

The $0.35\mu m$ CMOS technology was used as the reference technology. Equation 3.1 defines ΔT as the temperature difference between junction and the ambient. If ΔT is set

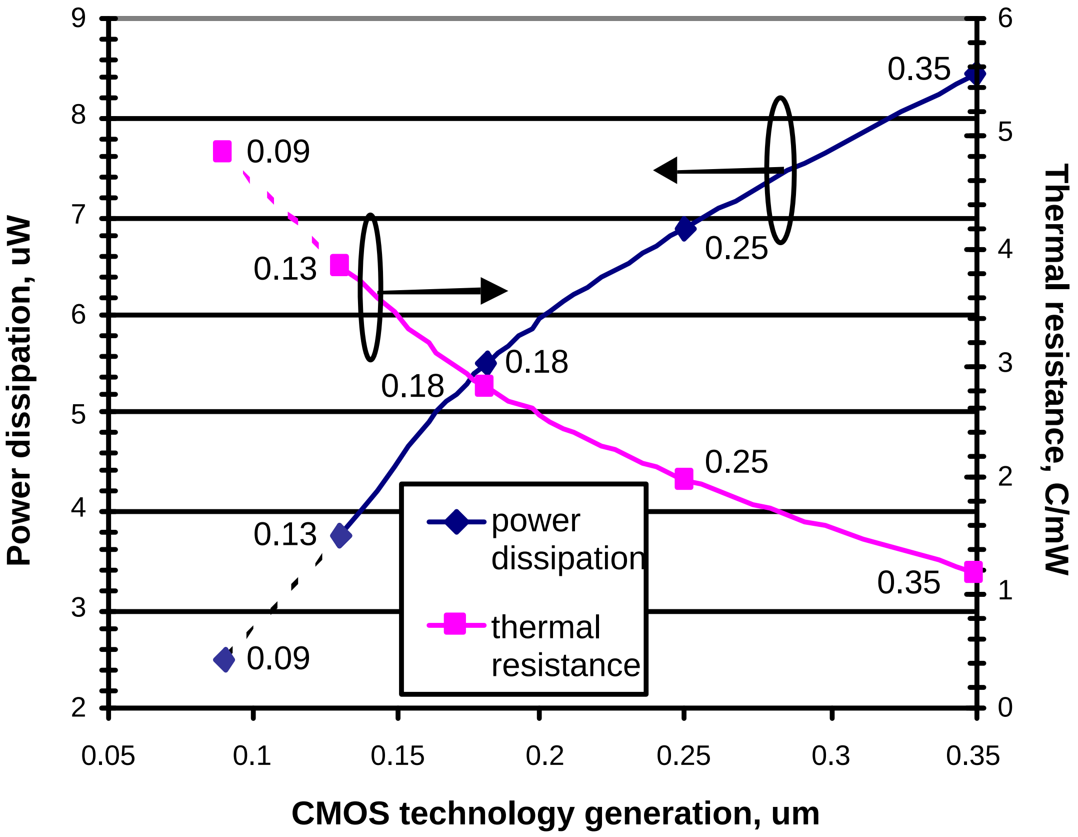


Figure 3.1: Inverter power dissipation and transistor thermal resistance for different CMOS technologies.

to unity for a $0.35\mu m$ technology, then the normalized change in ΔT with respect to the reference technology can be calculated. Using Equation 3.1 and data presented in Figure 3.1, the normalized average temperature increase for different technologies was estimated. For example, Equation 3.4 is used for calculation of $\Delta T_{0.25}/\Delta T_{0.35}$ ratio:

$$\frac{\Delta_{0.25}}{\Delta_{0.35}} = \frac{(T_j - T_a)_{0.25}}{(T_j - T_a)_{0.35}} = \frac{(P \times R_{ja})_{0.25}}{(P \times R_{ja})_{0.35}} \tag{3.4}$$

Figure 3.2 shows the normalized MOSFET junction temperature change with respect to the $0.35\mu m$ technology using Equation 3.4. As the technology shrinks from $0.35\mu m$ to $0.18\mu m$, the normalized temperature increased primarily from the increase in thermal resistance with scaling. However, scaling from $0.18\mu m$ to $0.09\mu m$ results in lower normalized MOSFET junction temperature with respect to $0.18\mu m$ technology. The reduction in normalized transistor temperature is due to the drastic reduction in power dissipation. The reduced parasitic capacitance is the primary reason for the reduced power dissipation. As a result of scaling from $0.18\mu m$ technology, P reduces faster than R_{ja} increases.

The increase in transistor density with scaling when estimating the average normalized temperature increase must also be considered. The density numbers were adopted from the International Technology Road map for Semiconductors (ITRS) [26][82]. Figure 3.3 shows the increased numbers of transistors and chip size with scaling. These graphs allow us to calculate the transistor density in the chip for the given technology.

The normalized temperature increase of a CMOS chip with technology scaling was calculated by multiplying the temperature increase per transistor in Figure 3.2 times the transistor density calculated from Figure 3.3. The results are shown in Figure 3.4. It can be concluded from Figure 3.4 that the normalized temperature increase of the chip is significantly elevated with CMOS technology scaling from 350 nm to 90 nm under normal operating conditions. The estimated junction temperature of a 90 nm CMOS chip is ~ 4.5 times higher than the junction temperature of a $0.35\mu m$ CMOS chip. This calculation assumed that the ambient temperature was the same for all analyzed technologies. The increase in chip junction temperature results in an exponential increase in cooling cost [33].

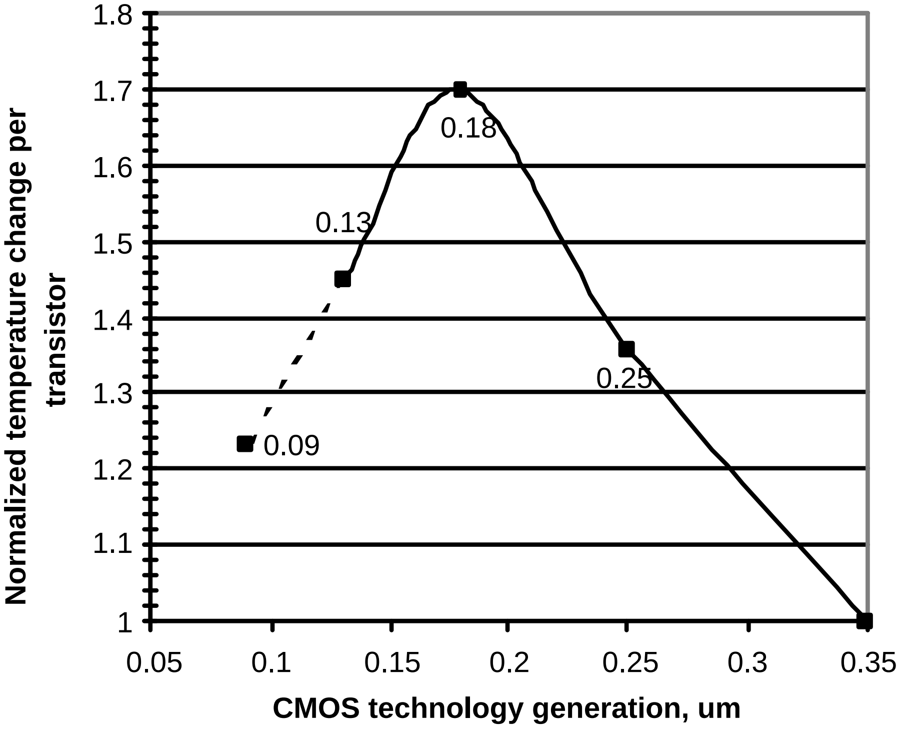


Figure 3.2: MOSFET junction temperature vs. technology

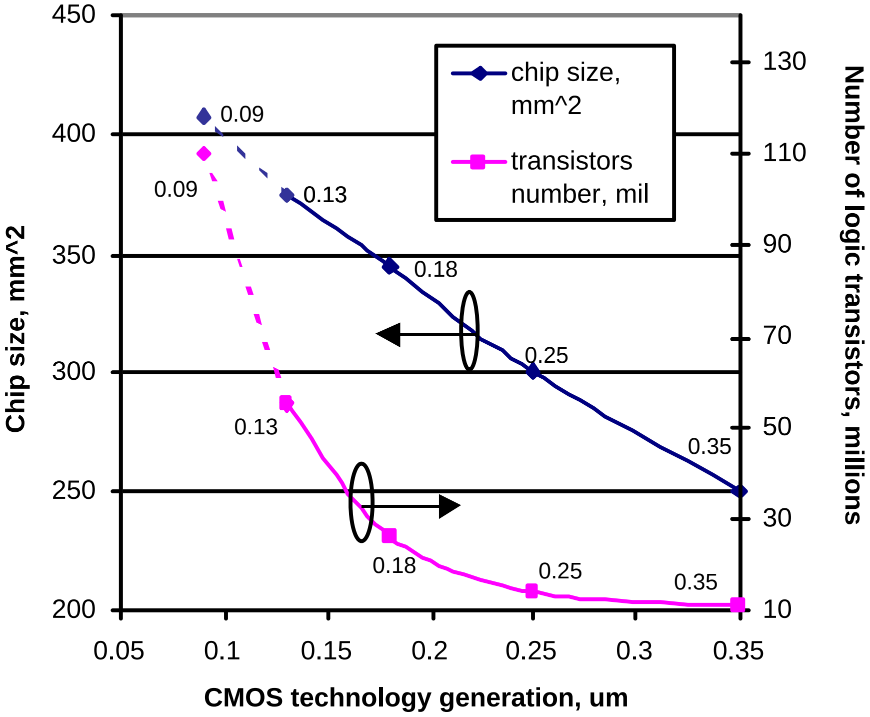


Figure 3.3: The trends of CMOS logic chips (data for graphs were adopted from [22,27]).

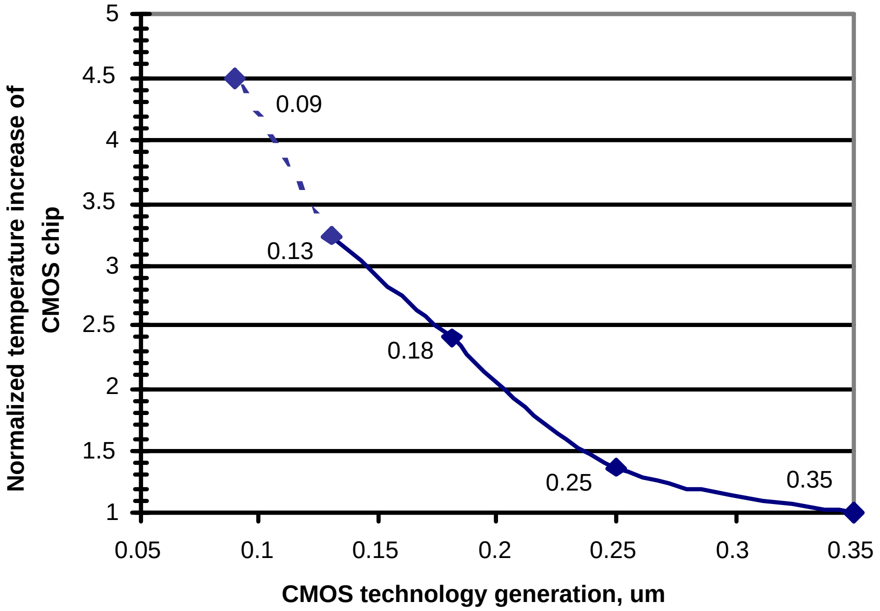


Figure 3.4: Normalized chip junction temperature increase with technology scaling for normal operating conditions.

3.2.2 Estimation of Junction Temperature Increase with Technology Scaling at Burn-in Conditions

The burn-in screening procedure weeds out latent defects from a product, and thereby improves the outgoing quality and reliability of the product. During burn-in, ICs are subjected to elevated temperature and voltage in excess of normal operating conditions for a specific period of time. This accelerates the product lifetime through the early part of its life cycle allowing removal of the products that would have failed during that time.

There are die level burn-in (DLBI) and wafer level burn-in (WLBI) techniques. DLBI can handle, contact, and do burn-in stress on several packaged die together, while WLBI has the ability to contact every die location and perform the burn-in test simultaneously on an entire wafer. For the DLBI, one must also consider the thermal impedance network of the package [47]. Once this network is known, then Equation 3.3 can be suitably modified to reflect the total thermal resistance (R_{ja}) of the die and many types of package. In this work, the focus was on the intrinsic behavior (junction temperature estimation) of the silicon die under burn-in conditions for the sake of simplicity. In other words, the thermal impedance network of the package is not considered.

The average inverter power for different operating conditions and technologies (Table 3.1) was estimated by simulating the inverters at different temperatures and V_{DD}. For burn-in, the stress temperature was varied from $25°C$ to $125°C$. Similarly, the stress voltage was varied from nominal V_{DD} for the given technology to $V_{DD} + 30\%$, and in this simulation (BSIM model level 49) the inverter input was grounded. The simulated I_{av} and the calculated values of P and ΔT are given in Table 3.2, where I_{av} and P are the average current and power dissipation of an inverter, and ΔT is $(T_j - T_a)$ per $1mm^2$ of chip area calculated using Equation 3.5.

$$\Delta T = P_{transistor} \times R_{ja-transistor} \times \frac{D_{density}}{2} \quad [\frac{^{\circ}C}{mm^2}] \tag{3.5}$$

where $P_{transistor}$ is the power dissipation of the off-mode transistor in the inverter, $R_{ja-transistor}$ is the thermal resistance of the on-transistor in the inverter, and $D_{density}$ is the transistor density in the CMOS chip. For a given technology, the thermal resistance was extracted from Figure 3.1 and the transistor density was calculated from Figure 3.3. It was assumed that the circuit under study is a fully static CMOS design. Therefore, half of the total transistors are in the off-mode during burn-in, and this was taken into account by dividing $D_{density}$ by 2 in Equation 3.5.

CMOS	$25°C$	$25°C$	$25°C$	$85°C$	$85°C$	$85°C$	$125°C$	$125°C$	$125°C$
Tech., V_{DD}	I_{av}, pA	P, pW	$\Delta T,^{\circ} C/mm^2$	I_{av}, nA	P, nW	$\Delta T,^{\circ} C/mm^2$	I_{av}, nA	P, nW	$\Delta T,^{\circ} C/mm^2$
0.35μm, 3.3 V	7.7	25	0.00071	0.07	0.23	0.0066	2.05	6.77	0.2
0.35μm, 3.8 V	9.2	35	0.00099	0.084	0.32	0.0091	2.15	8.17	0.23
0.35μm, 4.3 V	11.1	47.7	0.0014	0.11	0.47	0.014	2.27	9.76	0.28
0.25μm, 2.5 V	19.3	48.3	0.0023	0.418	1.04	0.05	3.96	9.9	0.29
0.25μm, 2.9 V	22	63.8	0.0031	0.047	1.36	0.065	4.41	12.80	0.35
0.25μm, 3.25 V	25	81.3	0.0039	0.531	1.75	0.08	4.81	15.87	0.45
0.18μm, 1.8 V	90.5	163	0.02	1.33	2.39	0.24	8.96	16.13	0.97
0.18μm, 2.1 V	101	210	0.022	1.48	3.08	0.31	9.75	20.48	1.23
0.18μm, 2.35 V	112	264	0.027	1.62	3.81	0.39	10.9	25.6	1.51
0.13μm, 1.2 V	766	920	0.2	8.45	10	2.32	28	34	7.79
0.13μm, 1.4 V	1200	1680	0.38	12.3	17	3.94	34	47	10.97
0.13μm, 1.56 V	1860	2900	0.67	17.7	27.6	6.4	55	85	19.81

Table 3.2: DC simulation (I_{av}) and calculation results (P, ΔT) of CMOS inverters for different technologies.

Since there was no access to industrial HSPICE device models for the 90 nm CMOS technology, the HSPICE simulations in Cadence for this technology generation could not be used. To predict the possible increase of average junction temperature in CMOS chips under burn-in conditions, an NMOSFET at stressed operating conditions was simulated using the 2-D device simulator ”Microtec” [20]. The MOSFET parameters used for device simulations are given in Table 3.3. The simulation results correspond to DC characteristics of 90 nm transistors [38][64][28], such as $V_T = 0.2 - 0.28V$, $I_{ON} = 600 - 750\mu A/\mu m$ and $I_{OFF} = 20 - 100nA/\mu m$. These devices were developed for ultra high performance applications (UHP). Low power (LP) medium speed [38][28] devices assume $V_{TH} = 0.3 - 0.35V$, $I_{ON} = 480 - 520\mu A/\mu m$ and $I_{OFF} = 0.18 - 0.5nA/\mu m$. High performance (HP) applications assume a leakage current of approximately $10nA/\mu m$ [81].

	UHP	**LP**
Subthreshold doping, cm^{-3} (p-type)	5×10^{15}	5×10^{15}
Sorce/Drain doping, cm^{-3} (n-type)	3×10^{20}	3×10^{20}
V_T adjusted doping, cm^{-3} (p-type)	1.8×10^{18}	3×10^{18}
Punch through doping, cm^{-3} (p-type)	5×10^{19}	8×10^{19}
Effective gate oxide thickness. $A°$	18	18
L_{eff}/W, $nm/\mu m$	63/2	63/2
Nominal $V_{DS} = V_{DD}$, V	1.0	1.0

Table 3.3: N-MOSFET parameters used for simulations.

In this section, UHP and LP devices were considered as worst and best cases with respect to power consumption during burn-in. The transistor parameters obtained from simulations under normal operating conditions are presented in Table 3.4. The dominant components of the leakage current in a sub-100 nm MOSFET are sub-threshold, band-to-band tunneling, and gate oxide tunneling currents [27].

	L_{eff}, nm	V_T, V	I_{ON}, $\mu A/mum$	I_{OFF}, nA/mum
UHP	63	0.25	600	30
LP	63	0.35	440	0.6

Table 3.4: DC parameters for an N-MOSFET emulated in 90-nm CMOS technology ($V_{DD} = 1V$, $T = 25°C$).

The simulation results of an averaged sized MOSFET ($W/L = 2.0\mu m/0.1\mu m$) under stressed operating conditions are given in Table 3.5. In this table, P is the power dissipation of an off-mode inverter transistor that was obtained from device simulations. ΔT (thermal density) is the $(T_j - T_a)$ per $1mm^2$ of CMOS chip that was calculated using Equation 3.5. The transistor density in a CMOS chip was assumed to be $0.27 millions/mm^2$ (Figure 3.3). When ΔT in Table 3.2 and Table 3.5 was calculated, it was assumed that each off-mode transistor in a $1mm^2$ chip area was an independent heat source. The total junction temperature increase of this area over ambient temperature was defined as the product of heat source density and the junction temperature increase of a single transistor. In practice, the thermal coupling effect of transistors on a chip must be considered, and this depends on layout. In the first order approximation, the thermal coupling effect of transistors was neglected in these analysis. Table 3.2 and Table 3.5 show that the average leakage current and dissipated power is increased by at least two orders of magnitude by technology scaling if the ambient temperature is $85°C$ or less. At $125°C$, the increase in current and power dissipation with technology scaling is relatively less. However, the increase in ΔT is more dramatic owing to increased transistor density, leakage current, and the thermal resistance.

The normalized temperature increase of a CMOS chip with scaling at burn-in conditions is shown in Figure 3.5. The plot with diamond symbols depicts the normalized T_j

	LP	LP	LP	UHP	UHP	UHP
$V_{DD}(V)$	1.3	1.15	1.0	1.3	1.15	1.0
$P(pW)$, $-100°C$	0.34	0.208	0.124	0.120	0.084	0.057
$\Delta T(°C/mm^2)$, $-100°C$	1.5×10^{-4}	9.1×10^{-5}	5.4×10^{-5}	0.052	0.036	0.021
$P(nW)$, $0°C$	0.75	0.51	0.32	44.2	29	18.4
$\Delta T(°C/mm^2)$, $0°C$	0.33	0.23	0.14	19.3	12.66	8.03
$P(nW)$, $25°C$	2.7	1.84	1.2	130	82.8	60
$\Delta T(°C/mm^2)$, $25°C$	1.18	0.81	0.53	56.8	36.15	26.2
$P(nW)$, $85°C$	21.23	14.63	9.8	770	506	328
$\Delta T(°C/mm^2)$, $85°C$	9.27	6.39	4.28	336.2	221	143.2
$P(nW)$, $125°C$	152.9	107.6	74.4	3084	2047	1344
$\Delta T(°C/mm^2)$, $125°C$	66.75	46.99	32.49	1346.6	893.8	586.8

Table 3.5: Predicted power dissipation and junction temperature increase in a CMOS inverter (90 nm CMOS technology).

increase if $T = 125°C$. For 90 nm technology the increase in T_j is different for the high performance or low power process. If all the transistors are implemented with low V_T UHP devices (unrealistic) then the normalized T_j is increased by approximately 5000X compared to $0.35\mu m$ CMOS. On the other hand, if all the transistors are implemented with LP devices, then the T_j is increased by approximately 230X. It should be noted that most of the transistors on a chip will be implemented with LP devices. However, if the T_a is reduced by $10°C$ for each technology generation, the normalized T_j is also reduced as shown by the plot with square symbols. Similarly, leakage reduction techniques can also be employed to further reduce the increased normalized temperature with scaling [16][41]. If such techniques are employed as well as the T_a being reduced by $10°C$ for each technology generation, the normalized T_j increase for 90 nm CMOS with respect to $0.35\mu m$ CMOS becomes relatively small (7-8X). In spite of the reduction in T_a and the use of leakage reduction techniques, the increase in T_j is still clearly unacceptable. Obviously, burn-in conditions should be carefully optimized for 130 nm and 90 nm CMOS technologies to

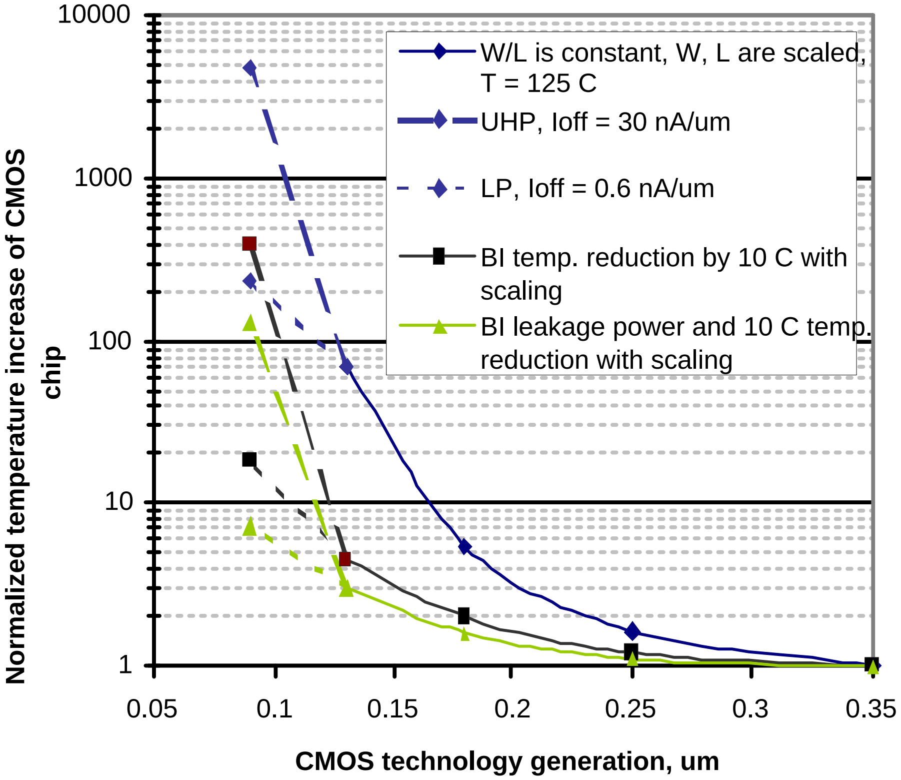


Figure 3.5: Normalized chip junction temperature at $V_{DD} + 30\%$ burn-in condition.

reduce the risk of chip over stressing during burn-in.

3.3 Impact of Package Thermal Resistance on Burn-in

High performance VLSI circuits, such as microprocessors, significantly challenge power delivery and heat removal due to smaller dimensions and increasing power dissipation. Technical challenges in the thermal management of microprocessors arise from two causes [52]:

- Increased dynamic and leakage power dissipation associated with technology scaling.
- Heat removal from localized hot spots.

The former is especially important for burn-in since the leakage power is exponentially increased under stress conditions. Typically, thermal management features are integrated in packages to spread heat from die to the heat sink. The heat sink dissipates the heat into local environments. A typical thermal resistance network of a packaged die is shown in Figure 3.6. By definition, the case temperature (T_c) is the temperature at the external surface of the package. All semiconductor packages have multiple elements. In the simplest form these elements include the semiconductor die, thermal interface material, and the heat sink base. The thermal conductivity of these package elements for the Pentium III Xeon microprocessor is given in Table 2. In a common case, the junction temperature increase over ambient temperature has three components [91]:

$$\Delta T = P * [R_{th}(Die - pack) + R_{th}(Pack - sink) + R_{th}(Sink - amb)] \tag{3.6}$$

where $R_{th}(Die-pack)$, $R_{th}(Pack-sink)$, $R_{th}(Sink-amb)$ are the die to package, package to heat sink, and heat sink to ambient thermal resistances, respectively, and P is the

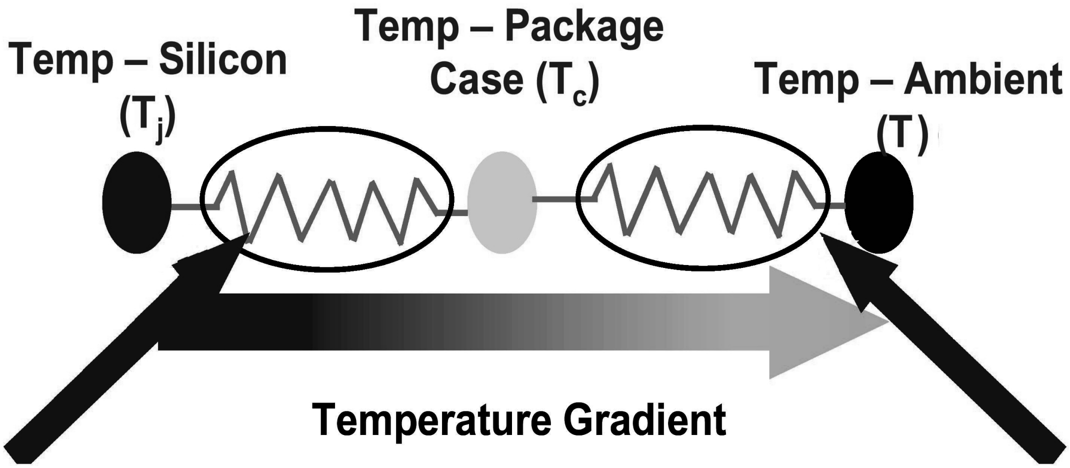


Figure 3.6: Thermal resistance network of a packaged die: (1) junction to case (package), (2) case to ambient (heat sink) [52].

Packaged component	**Conductivity, W/mK**
Silicon die	120
Thermal interface material	3.8
Heat sink base	180

Table 3.6: Thermal conductivity of package components [32].

total power dissipation of the chip. The first component in Equation 3.6 is discussed in previous sections. The third component is determined by the cooling techniques and will be considered in the next section. Here, the second component in Equation 3.6 is considered and can be rewritten as follows:

$$\Delta T(Package) = P_{MOSFET}.\frac{D}{2}.R_{th}(Pack - sink) \tag{3.7}$$

where P_{MOSFET} is the transistor power dissipation, and D is the transistor density. The package to heat sink thermal resistance, $R_{th}(pack - sink)$, is crucial to removing heat during burn-in. Values of $0.9 - 1.2°C/W$ were reported for $R_{th_{PH}}$ in 350 nm technology [2][1]. It is predicted that a reduction of approximately 22% in $R_{th}(pack - sink)$ per technology generation is required to just compensate for the increased power density with technology scaling [7]. Figure 3.7 shows these projections for the 350 nm technology to 90 nm technology.

3.4 Cooling Techniques for Burn-in

Low power devices can be burnt-in without attention to thermal considerations. However, as power dissipation increases with technology scaling for high performance chips, burn-in requires advanced cooling concepts and additional hardware to facilitate direct contact between the heat sink and the die. Advanced burn-in ovens should provide uniform temperature distribution in the chamber and precise temperature control for each individual device. The power dissipation within one lot of devices can vary by 40% due to manufacturing variations and different test vectors applied during burn-in. This variation in power, and approximately 30% variation in oven airflow, can create a significant variation in package temperature [34]. If the device becomes too hot, it may be damaged while other devices may not be adequately burnt-in. To uniformly stress all devices, each

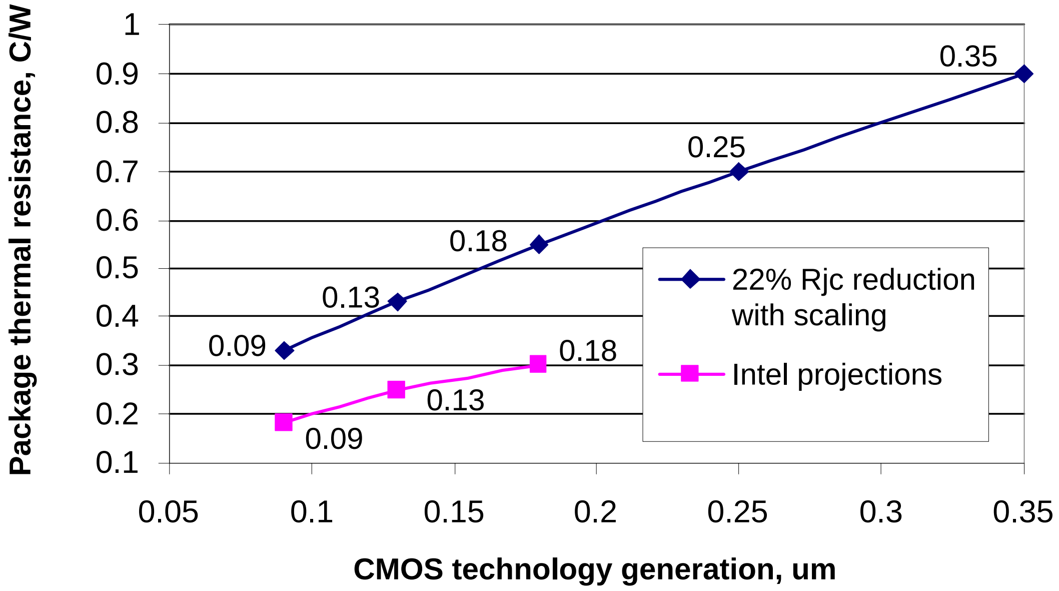


Figure 3.7: Reduction of package thermal resistance with technology scaling [7].

package device temperature must be kept close to the specified burn-in temperature. This is achieved by developing advanced cooling techniques and burn-in boards with embedded thermal sensors.

3.4.1 Power Limitations of Burn-in Equipment

The total number of die that can be simultaneously powered-up for burn-in testing will likely be limited by the maximum power dissipation capacity of the burn-in oven. A typical oven may contain several hundred dies. If all dies are active, then the total power dissipation can reach the several kilowatt range. Typically, burn-in ovens have a maximum dissipation power between 2500-6500 Watts [39]. The power dissipation of a single transistor in an inverter in static stressed conditions and the number of transistors of the logic chip can be used to estimate different CMOS technologies. Then the maximum number of die for different technologies that can be simultaneously powered in a burn-in oven can be estimated using Equation 3.8:

$$N_{dies} = \frac{P_{oven}}{P_{transistor} * \frac{N_{transistors}}{2}} \tag{3.8}$$

where P_{oven} is the maximum power dissipation of the burn-in oven at stressed conditions, $P_{transistor}$ is the power dissipation of a single transistor at static stressed conditions for the given technology, and $N_{transistors}$ is the total number of transistors in the logic chip for the given technology. Equation 3.8 assumes that 50% of the total number of transistors are off at any point during burn-in assuming fully static CMOS design. Results are shown in Figure 3.8. Burn-in ovens, such as the PBC1-80 of Dispatch Industries [39] and Max-4 of Aehr Test Systems [76] have maximum power dissipation of about 2500 and 15,000 watts, respectively, at $125°C$. The room ambient temperature is assumed to be $25°C$.

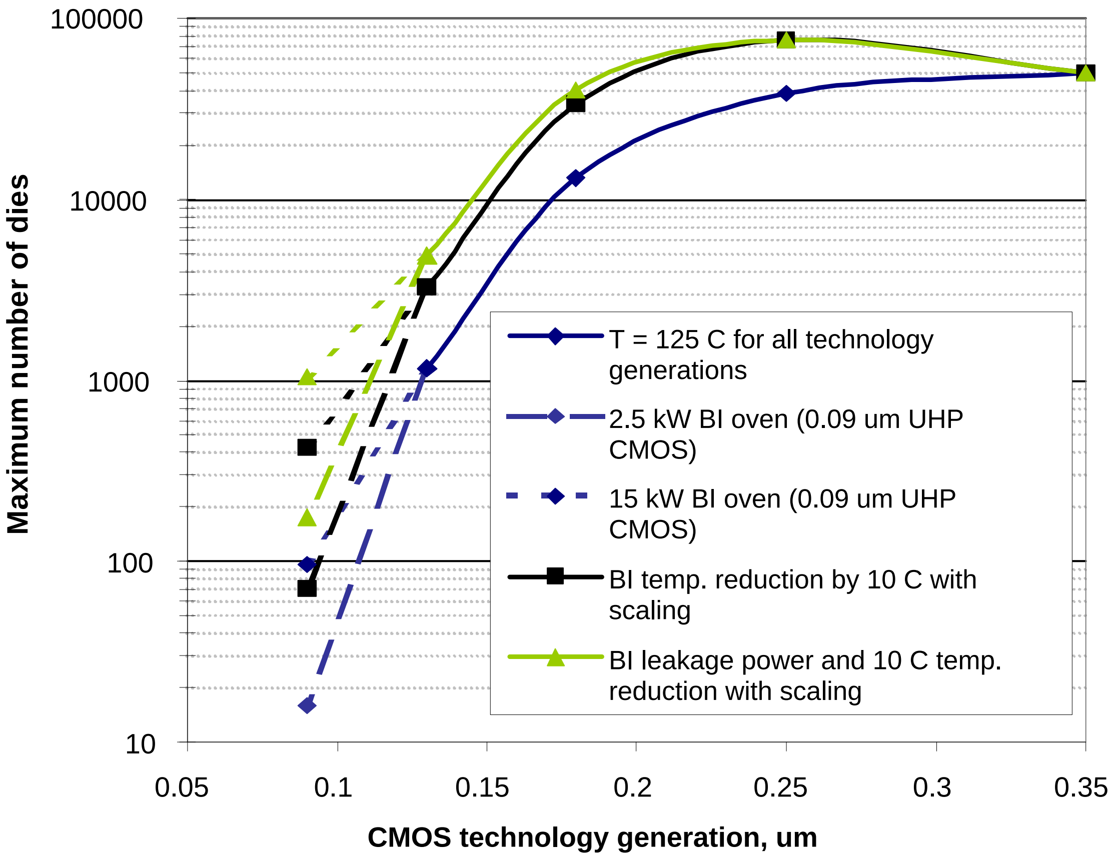


Figure 3.8: Maximum number of dies for one burn-in load with scaling.

3.4.2 Air Cooling Technique

For CMOS IC technologies of $0.35\mu m$ and above, generally IC junction heating during burn-in has not been a major issue and the oven temperature could be easily set to avoid temperature-related over stress. However, for 0.25 um technology and below, device self-heating has been described to become a more significant issue and air-cooling techniques began to be implemented to remove heat from each device and the oven.

Air-cooled burn-in ovens are reasonably effective in heat removal from devices dissipating up to 30-40 watts [34]. Often, an air-cooled heat sink and embedded thermal sensors are used to control the individual temperature of each device. The air temperature and air velocity are dependent on the device power, the overall thermal resistance of the heat sink assembly and burn-in socket, and the required package temperature. The air temperature and velocity must be controlled so that the embedded heat sink can limit the device temperature increase over the range of heat dissipation. The device temperature can be controlled in the range of $50°C$-$150°C$ with an accuracy of $3°C$ [34]. Device temperature is usually measured by attaching a small thermocouple directly on the device or by using sensors integrated into the device [78].

Another air-cooling technique was developed for device power dissipation from 35 to 75 watts [34]. This approach uses a small fan mounted above the heat sink of each device. The amount of allowable device power dissipation is a function of the air temperature, air velocity, thermal resistance of the heat sink, and the package.

To ensure quality output, ovens are designed to ensure that the temperature distribution across all the boards is uniform and adequate. The level and uniformity of the temperature across the burn-in boards is controlled by the total airflow induced in the oven and the uniformity of the airflow distribution between the boards. The design of an airflow network becomes increasingly more complicated as device power dissipation increases [49].

3.4.3 Liquid Cooling Technique

As power dissipation increases beyond 75 watts per device, the thermal resistance of the package to ambient must be lowered to allow removal of excess heat. Air-cooling burn-in techniques are not effective for power dissipation in this range and it has fostered the development of liquid-based cooling techniques. Figure 3.9 illustrates one such technique [34]. A temperature sensor embedded in the heat sink measures the device temperature. Helium is injected into the heat sink to provide a lower thermal interface between the device and the heat sink. This technique lowers the heat sink to ambient thermal resistance by approximately 40%.

Each heat sink has a temperature-controlled heater. The burn-in ovens with liquid-cooled heat sinks can burn-in devices that dissipate over 150 W of power [29][65]. In such ovens, the ambient temperature for each device can be optimized for optimal burn-in conditions. This is important since self-heating dissipation can vary significantly due to inherent process spreads in scaled technologies. The thermal control during test and burn-in of devices with high leakage power dissipation (above 75 Watts) plays a key role in increasing the post burn-in yield. Special thermal test chips and modules were developed to measure temperature gradients in packages and heat sinks in burn-in equipment [29][65]. For example, IBM used a TV994 thermal test chip for burn-in equipment qualification. This 14.7 mm^2 chip has nine small resistive temperature detectors (RTD) and four large heater resistors, one covering each quadrant of the chip [29]. The thermal interface tests evaluate temperature gradients within the device and between the device and heat sink. Temperature differences are normalized with respect to applied device power. The test is used to optimize and evaluate factors such as heat sink material, flatness and various properties of interface pads, and liquids and gases that can be between the chip and heat sink.

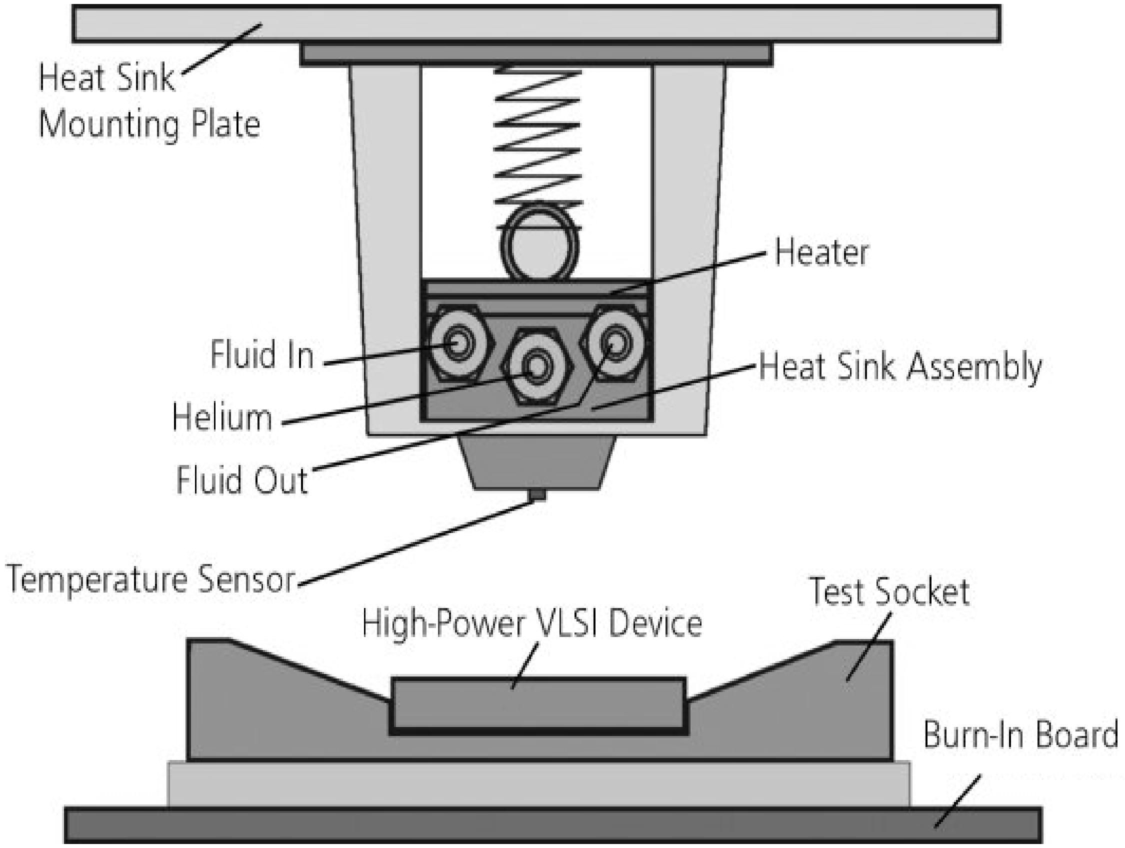


Figure 3.9: Water-cooled heat sink, adopted from [34].

3.5 Burn-in Limitations and Optimization with Respect to Yield and Reliability Issues

Yield and reliability are two important factors in semiconductor manufacturing. Typically three parameters significantly affect the yield and reliability of ICs [45]:

- Design-related parameters (chip area and gate oxide thickness).
- Process-related parameters (defect distribution and density).
- Operation-related parameters (voltage and temperature).

It has been experimentally verified that defects that cause burn-in failures (early-life reliability failures) are fundamentally the same in nature as defects that cause wafer probe failures (yield failures) [10][24].

Researchers have also identified two key reliability indicators in order to optimize yield during burn-in:

- Local region yield.
- The number of defects that have been repaired (for chips containing redundancy).

Experimentally, it has been shown that die with many faulty neighbors can pose a significantly greater early-reliability risk than chips with few faulty neighbors [9]. An IC with a redundancy-related repair is more likely to have a latent defect mechanism resulting in early life failure [10].

The key to optimizing burn-in lies in identifying those die that most likely to fail during burn-in before the burn-in is actually performed. Once identified, die of higher reliability risk may be subjected to more rigorous testing (longer burn-in duration), while those dies deemed more reliable may have a reduced stress, or no stress at all. Barnett et al. proposed the post burn-in yield model, which include the burn-in time as a parameter [9]. It was assumed that the average number of latent defects (λ_L) per chip is time-dependent as follows:

$$\Lambda_L(t) = \alpha.\gamma.(1 - Y_K^{1/\alpha}).(\frac{t}{\tau})^{\beta} \tag{3.9}$$

where α is the defect clustering parameter, $\gamma = 0.01 - 0.02$ is the fitting parameter, Y_K is the wafer test yield (yield before burn-in), t is the burn-in time in hours, and β is the shape parameter of the Weibull distribution of the reliability function. The post burn-in reliability yield (i.e. the number of dies surviving burn-in) is modeled as follows:

$$R(t) = [1 + \frac{\lambda_L(t)}{\alpha}]^{-\alpha} \tag{3.10}$$

Kim et al. [46] developed another model for post burn-in reliability (R) and yield loss (Y_{loss}), which will be discussed later in this section.

Burn-in removes the infant mortality device population hence improving the outgoing device reliability. However, burn-in may affect the post burn-in yield of ICs since latent defects may become enhanced during burn-in, with a resultant increase in post burn-in yield loss. The amount of yield loss depends on burn-in conditions (voltage, temperature, time). Since the stress voltage and the stress temperature provide the acceleration during burn-in, the burn-in time is the parameter that is manipulated to control the post burn-in yield loss using above mentioned models. In practice, many IC manufactures reduce the burn-in time to 10 hours or even skip burn-in, when the yield before burn-in is high ($\sim$ 98%) and burn-in escapes is low ($\sim$ 100 PPM) [46]. The amount of burn-in escape is estimated by the early failure rate test, which is performed on 10000 final products from at least three lots with duration approximately 12-48 hours under burn-in conditions.

Several reliability failure mechanisms are accelerated by temperature, so burn-in testing is done at elevated temperature. These mechanisms include metal stress voiding and electromigration, metal slivers bridging shorts, contamination, and gate-oxide wear out and breakdown [68]. However, there are physical and burn-in equipment related limitations for temperature and voltage stress. Die failure rate (failures per million) increases exponentially with temperature for most failure mechanisms [23]. As a result, the yield loss may increase if the burn-in conditions are over stressed. Hence, we should optimize the junction temperature of die for normal and burn-in conditions.

3.5.1 Physical and Practical Limits of Junction Temperature

The maximum operating temperatures for semiconductor devices can be estimated from the semiconductor intrinsic carrier density that depends on the band-gap of the material. When the intrinsic carrier density reaches the doping level of the active region of devices, electrical parameters are expected to change drastically. The highest reported operating junction temperature is about $200°C$ in standard silicon technology [90]. At this temperature, the circuit performance is reduced substantially. The temperature will affect thermal conductivity, built-in potential, threshold voltage, and *pn* junction reverse current. Several practical considerations limit the junction temperature to a much lower value. A value of $150°C$ for junction temperature is often used for ICs as the limit [26].

The peak junction temperature in a PowerPC microprocessor implemented in a $0.35\mu m$ CMOS technology with a $0.3\mu m$ effective transistor channel lengths is about $90°C$-$100°C$ at an operating speed of 200-250 MHz [73][67]. If this is used as the reference temperature and assuming that Figure 3.4 estimates the junction temperature increase with reasonable accuracy and package thermal resistance remains the same, then one can expect a 2.4X increase in junction temperature for the same microprocessor implemented in a $0.18\mu m$ CMOS technology. Hence, the die junction temperature should be approximately $156°C$-$180°C$.

These values are obtained assuming cooling, packaging and circuit techniques remain the same when moving from $0.35\mu m$ technology to $0.18\mu m$ m technology. However, improved cooling and packaging considerations will reduce the temperature to a much lower value. Similarly, circuit techniques such as transistor stacking, dual-threshold transistors, reverse body bias, etc. can reduce substantially leakage current and the junction temperature.

3.5.2 Optimization of Burn-In Stress Conditions with Technology Scaling for Constant Reliability

The optimal burn-in conditions for maintaining the projected failure rate require that the defect distribution models and their growth models be studied. The post burn-in reliability (R) and yield loss (Y_{loss}) have been studied [83][46]. T. Kim, et al., [46] proposed the following models for post burn-in reliability and yield loss shown in Equation 3.11 and Equation 3.12.

$$Y_{loss} = Y.(1 - Y^{\frac{\nu}{1-\nu}}) \tag{3.11}$$

$$R = Y^{\frac{1}{(1-u)^2-1}} \tag{3.12}$$

where Y is the yield before burn-in, and u, ν are constants that depends on the stress temperature and voltage. Using the $1/E$ gate oxide breakdown model and the post burn-in yield loss model, Vassighi, et al., demonstrated that the post burn-in yield loss increases exponentially with increasing stress temperature for a given stress voltage [83]. This result was obtained for a $0.18\mu m$ CMOS technology ($T_{OX} \gg 41A°$).

Hence, over stressing a die during burn-in may significantly reduce the post burn-in reliability and increase the yield loss, especially when the junction temperature at burn-in and normal operating conditions are increased with technology scaling. Thus, to a first order, we want a constant reliability during burn-in with technology scaling. The burn-in temperature and voltage should be optimized for different CMOS technologies to maintain the average junction temperature of the die at the same fixed level. If electrical defect densities are equal, then we assume that the post burn-in reliability for an advanced CMOS technology should not be worse than the post burn-in reliability for the $0.35\mu m$ CMOS technology. This means that the junction temperature increase over ambient temperature during burn-in for subsequent technologies should not be higher than the burn-in junction temperature increase for $0.35\mu m$ CMOS technology. Table 3.2 shows that for $0.35\mu m$ CMOS technology, the junction temperature increase (ΔT) over ambient stressed temperature per mm^2 of chip is $0.28°C$ at $V_{DD} = 4.3V$, $T = 125°C$. The horizontal line on Figure 3.10 illustrates this limit. Now for $0.25\mu m$ technology, if the $\Delta T/mm^2$ versus stress temperature is plotted for three different stress voltages, it results in three different curves. Subsequently, the optimal burn-in temperature where the horizontal line ($\Delta T = 0.28°C/mm^2$) intersects with graphs can be found. Similarly, the optimal burn-in temperature for other technologies can be found using data from Table 3.2 and Table 3.5. The results are shown in Figure 3.11 where the optimal burn-in temperature is presented for different technologies. Squares represent the data points for each technology. In this figure, the stress voltage is kept at $V_{DD} + 30\%$ for each technology. These data points were plotted ensuring that the average junction temperature increase over ambient (ΔT) for die in these technologies is the same as the average ΔT increase for a $0.35\mu m$ CMOS technology. Hence, it is expected that the post burn-in reliability for scaled CMOS technologies has the same value as the post burn-in reliability for $0.35\mu m$ CMOS technology.

Figure 3.11 shows that the optimal burn-in temperature is reduced with scaling. This observation is in line with the recently presented data given for a $0.18\mu m$ microprocessor, where the burn-in temperature is $85°C - 90°C$ [54]. As mentioned before, if leakage

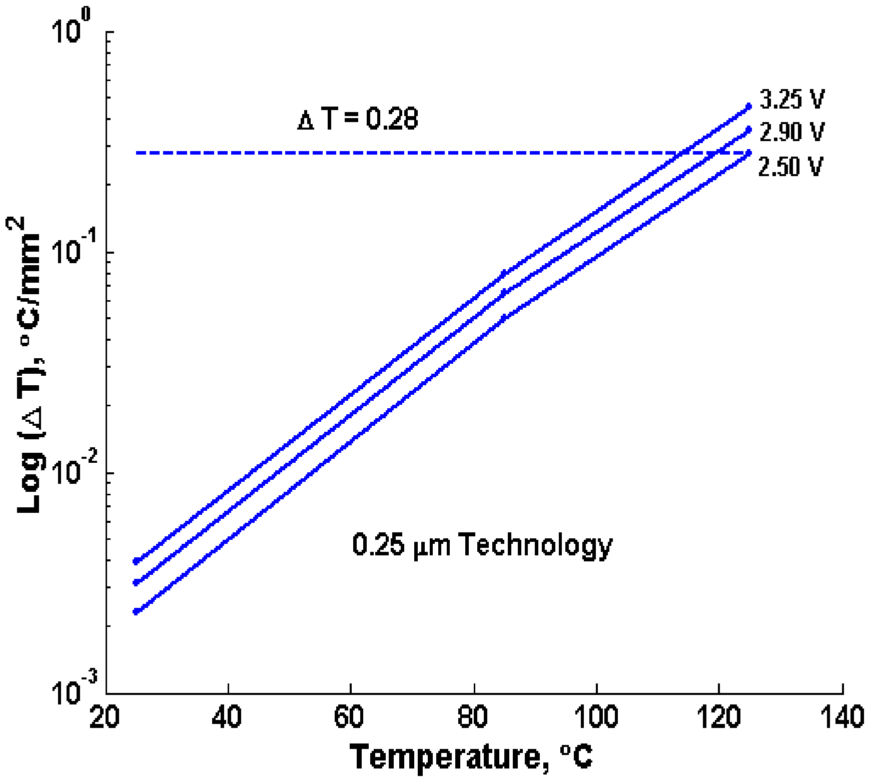


Figure 3.10: ΔT as a function of ambient temperature and V_{DD} for $0.25\mu m$ technology.

reduction techniques are employed (diamond data points), the optimal burn-in temperature is increased for $0.18\mu m$ or lower geometries. For example, according to this research, the optimal temperature for 130 nm technology ($V_{DD} \gg 1.4V$) is approximately $10°C$ (without leakage reduction techniques) and $35°C$ (with leakage reduction techniques).

Furthermore, if such a trend continues, we will have to cool future generations of CMOS devices during burn-in below room temperature, if we do not want the post burn-in reliability worse than that of the $0.35\mu m$ CMOS technology. For example, the estimated burn-in temperature for a 90 nm CMOS technology may be approximately $0°C$ to $15°C$. Note, that many future chips will use a mixture of technologies: UHP logic is for critical delay paths and LP logic is for the low activity SRAM cells [26].

3.6 Conclusion

The impact of technology scaling on the burn-in environment was investigated. The following conclusions are obtained: Firstly, there is a steady increase in the junction temperature with scaling. Under normal operating conditions, the normalized increase in junction temperature is estimated to be 1.45X/generation. Similarly, the normalized junction temperature increase under burn-in conditions becomes exponential with technology scaling if no leakage reduction techniques are used. On the other hand, if leakage reduction techniques are used, then an approximately linear increase in junction temperature can be obtained. As a consequence, the burn-in temperature must be reduced with scaling. Second, the number of dies that can be simultaneously burnt-in is reduced with technology scaling, because of the maximum power dissipation limit of presently available burn-in ovens. Finally, the optimal stressed temperature in a burn-in environment is significantly reduced with technology scaling.

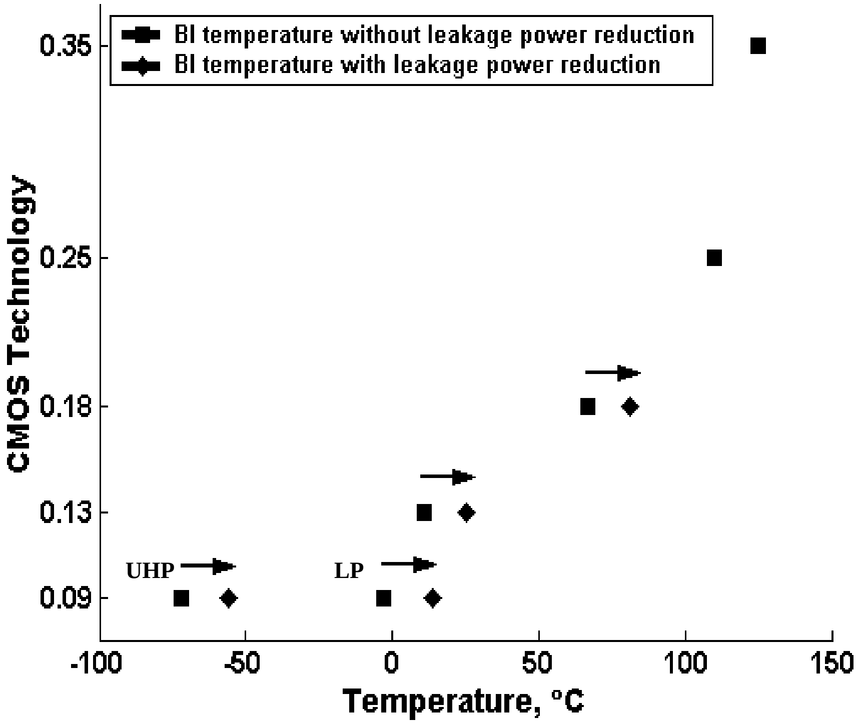


Figure 3.11: Optimal burn-in temperature for constant reliability.

It was also argued that deep sub-micron devices will require advanced packaging and liquid cooling techniques to lower the junction to ambient thermal resistance.

In scaled technologies, burn-in optimization for yield and reliability will be of crucial significance owing to larger number of design and technology variables. In some situations, individual chip level burn-in optimization will be necessary in order to provide optimum burn-in environment for each chip.

Chapter 4

Thermal Runaway Avoidance During Burn-in

The maximum operating temperatures for semiconductor devices can be estimated from semiconductor intrinsic carrier density, which depends on the band-gap of the material. When the intrinsic carrier density reaches the doping level of the active region of devices, then the electrical parameters change drastically. The highest operating junction temperature for standard silicon technology is about $200°C$, however the circuit performance is reduced substantially [90]. The influence of temperature on some important MOSFET parameters is summarized in Table 4.1. Several practical considerations limit the junction temperature to a much lower value. A limit of $150°C$ for junction temperature is often used for VLSI ICs [26]. The peak junction temperature of a PowerPC microprocessor implemented in a 350 nm CMOS technology was reported to be approximately $90°C$-$100°C$ at an operating speed of 200-250 MHz [73][30].

Lowering V_T to achieve higher performance leads to higher off-state leakage current, and this is the major problem facing burn-in and scaled nanometer technologies.

Parameter	Temperature dependence	Affected property
Thermal conductivity, K	$= T^{-1.6}$	Self heating
Built-in potential, V_{bi}	$\frac{KT}{q} ln(\frac{N_A N_D}{n_i T^2})$	+20% per 100K
Threshold voltage, V_T	$2si_B(T) + (4\epsilon_{Si} q N_A si_B(T)/C_I)^{0.5}$	-0.8 mV/K
pn junction reverse current	$a \times n_i^2(T) + b \times n_i(T)/\tau_{sc}$	10^2 to 10^4 per 100K

Table 4.1: Temperature-dependence of important Si-MOSFET parameters, data adopted from [90].

The total power consumption of a high performance microprocessor increases with scaling. Considering the increasing percentage of off-state leakage current at the 130 nm and sub-100 nm nodes under nominal conditions, the ratio of leakage to active power becomes adverse under burn-in conditions. Typically, clock frequencies are kept in the tens of MHz range during burn-in, which results in substantial reduction in active power. On the other hand, the voltage and temperature stresses cause the off state leakage power to be the dominant power component. Stressing during burn-in accelerates the defect mechanisms responsible for early life failures. Thermal and voltage stresses increase the junction temperature resulting in accelerated aging. Elevated junction temperature, in turn, causes leakage to further increase. In many situations, this may produce positive feedback leading to thermal runaway. Such situations are more likely to occur as technology is scaled to the nano-meter regime. Thermal runaway increases the post burn-in yield loss dramatically. Figure 4.1 shows a chip that has gone into thermal runaway. To avoid thermal runaway, it is crucial to understand and predict the junction temperature under stress conditions. Junction temperature, in turn, is a function of ambient temperature, package to ambient thermal resistance, package thermal resistance, and static power dissipation. Considering these parameters, one can optimize the burn-in environment to minimize the

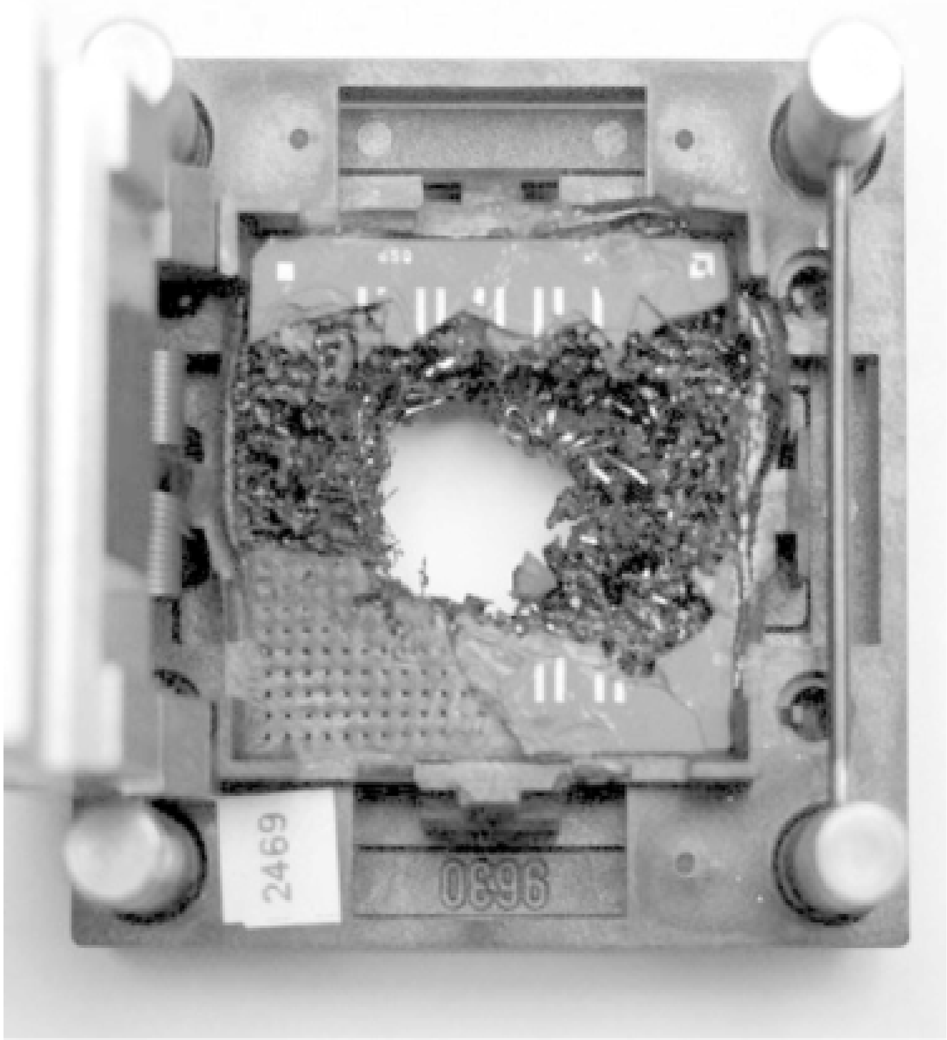


Figure 4.1: Test socket destroyed by thermal runaway [6].

probability of thermal runaway while maintaining the effectiveness of the burn-in test.

4.1 Junction Temperature Estimation Procedure

Historically, the burn-in environment temperature and voltage have been $125°C$ and $V_{DD}+30\%$ to $V_{DD}+40\%$, respectively. At the time, the leakage power was a non-issue. However, in sub-180 nm technologies, leakage power is significantly higher under burn-in conditions. Figure 4.2 shows the transistor leakage current increase for 130 nm CMOS technology at burn-in conditions. The figure shows the increase in leakage power with increasing temperature and voltage. As it can be seen from the graph, the leakage is increased by approximately 3-4X, going from nominal to burn-in conditions.

The junction temperature (T_j) of an IC is defined as the average temperature of the silicon substrate. T_j is a crucial parameter in reliability-prediction procedures and burn-in testing. Under burn-in conditions, accurate junction temperature estimation may reduce the thermal runaway probability since the margin between optimal burn-in conditions and thermal runaway is reduced as the technology is scaled. The junction temperature or T_j, is defined as [79]:

$$T_j = T_a + P \times R_{ja} \tag{4.1}$$

where T_a is the ambient or set point temperature, P is the device total power, and R_{ja} (R_{th}) is the junction-to-ambient thermal resistance. The power dissipation can be subdivided into dynamic and leakage components, as:

$$P = P_{Dynamic} + P_{Leakage} \tag{4.2}$$

$$P_{Leakage} = V_{DD} \times I_{Leakage} \tag{4.3}$$

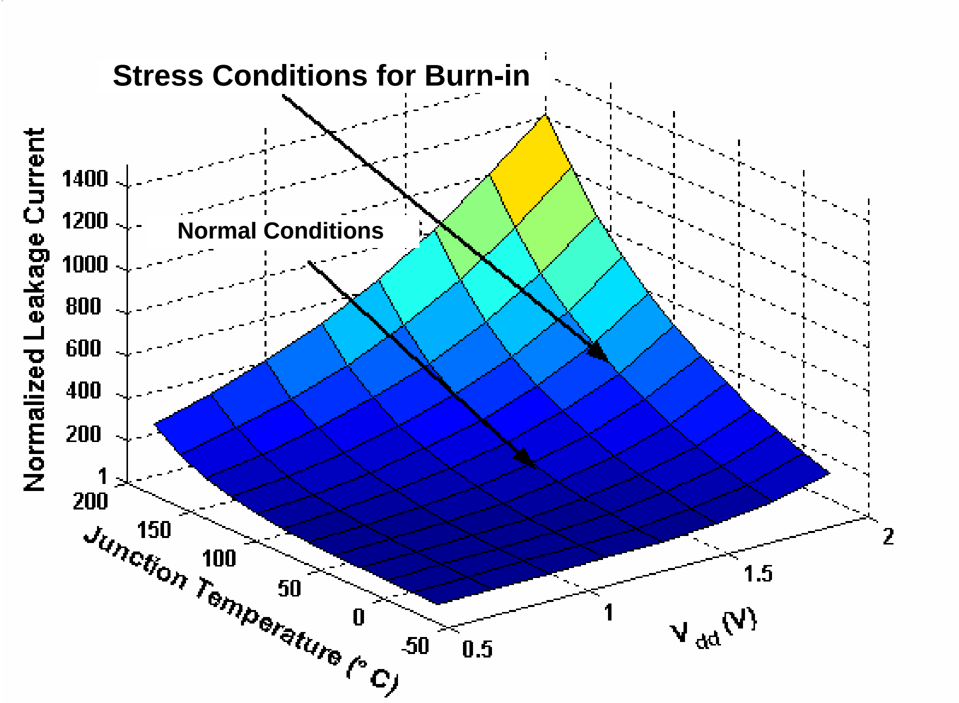


Figure 4.2: Off current of a NMOS transistor in terms of voltage and temperature for 130 nm CMOS technology (normalized to the off current for $V_{DD} = 0.6V$ and $-25°C$)

$$P_{Dynamic} = C \times V_{DD}^2 \times f_{toggle} \tag{4.4}$$

In equation 4.4, C is the total IC switching capacitance and f_{toggle} is the frequency that is used for node toggling during burn-in and can be expressed as:

$$f_{toggle} = \frac{I_{on}}{C_{gate} \times V_{DD} \times N} \tag{4.5}$$

where C_{gate} is the gate capacitance of a single gate and N is the number of logic stages in the critical path. To evaluate the junction temperature, T_j, under different environmental conditions, a program and a methodology has been developed [42]. Figure 4.3 depicts the flow chart of the program. At any initial temperature, the program reads the input current for a single transistor. Based on the circuit implementation and architecture, the total power is computed using Equations 4.3, 4.4, and 4.5 and the junction temperature is updated in Equation 4.1. Using this procedure [42], for any given voltage and process technology, the junction temperature is calculated and convergence of the obtained temperature is tested [85]. Depending on the result after several iterations the junction temperature will either converge to a temperature or will increase and lead the chip into thermal runaway.

4.2 Simulation Results

A 32-bit microprocessor in 130 nm dual-V_T CMOS technology was used to verify the procedure. The parameters of this program were calibrated to the experimental data from the microprocessor. These parameters include the burn-in stress voltage and temperature, I_{on} and I_{off} of the CMOS transistors, and the layout of the chip. Figure 4.4 depicts the electro-thermal simulation results for this microprocessor in the burn-in conditions. As illustrated in Figure 4.4, for air-cooled ovens if the ambient temperature is kept above $10°C$

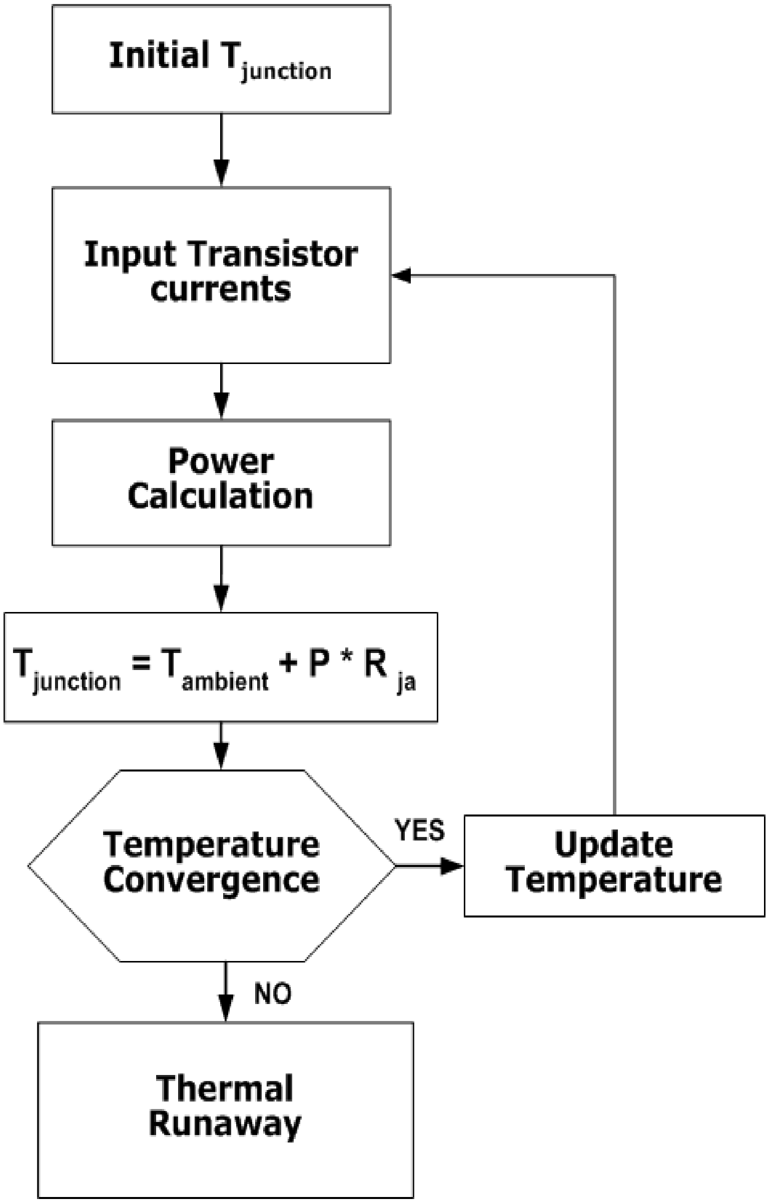


Figure 4.3: A procedure for junction temperature estimation [42].

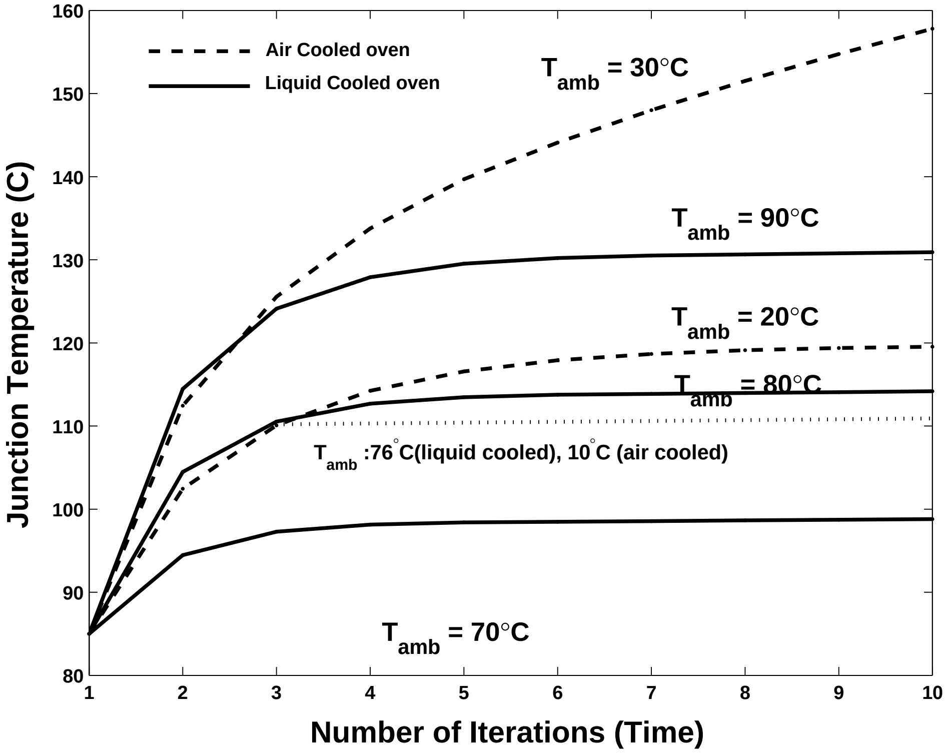


Figure 4.4: Steady-state junction temperature for 130 nm high performance microprocessor at burn-in conditions ($T_j = 110°C$ and $V_{DD} = 1.8V$).

the junction temperature starts rising and does not stabilize. This rise in temperature will lead the chip to thermal runaway. The same chip in a liquid cooled burn-in oven (solid lines) will tolerate up to $76°C$ of ambient temperature, which results in $110°C$ of junction temperature. Liquid cooled burn-in ovens with a junction to ambient thermal resistance of $0.5°C/W$ are able to transfer more heat from the chip than air cooled burn-in ovens with junction to ambient thermal resistance of $1.5°C/W$. Since the total power at burn-in condition $(T_j = 110°C, V_{DD} = 1.8V)$ for this chip is 66W, $1°C/W$ reduction in thermal resistance will allow us to perform burn-in with $66°C$ higher ambient temperature. The results in Figure 4.4 confirm that the ambient temperature is increased from $10°C$ to $76°C$ in liquid cooled ovens. It should be noted that since the ambient temperature in an air-cooled ovens cannot be less than the a room temperature, it is impractical to burn-in this microprocessor in air-cooled burn-in oven as at room temperature ambient, the chip will eventually go into thermal runaway.

The processors in a production line often have a skewed normal leakage distribution. The processors with larger off state leakage are more susceptible to thermal runaway. Since processors with higher leakage are also faster, the economic cost of losing them to thermal runaway is even higher than the processors with average leakage. Therefore, a flexible burn-in procedure must be tailored according to the leakage. The processors are categorized based on their leakage. Subsequently the burn-in procedure for each category is optimized to minimize the thermal runaway probability. The variations in leakage power are mostly due to process variations. Simulations were carried with 10% reduction in the channel length of the transistors in 130 nm technology. This resulted in a 3X increase in the sub-threshold current. This increase in the sub-threshold current increases the leakage power of the test chip under burn-in condition by 3X. This extra leakage increases the junction temperature. Figure 4.5 illustrates the simulation results of the chip that its transistors

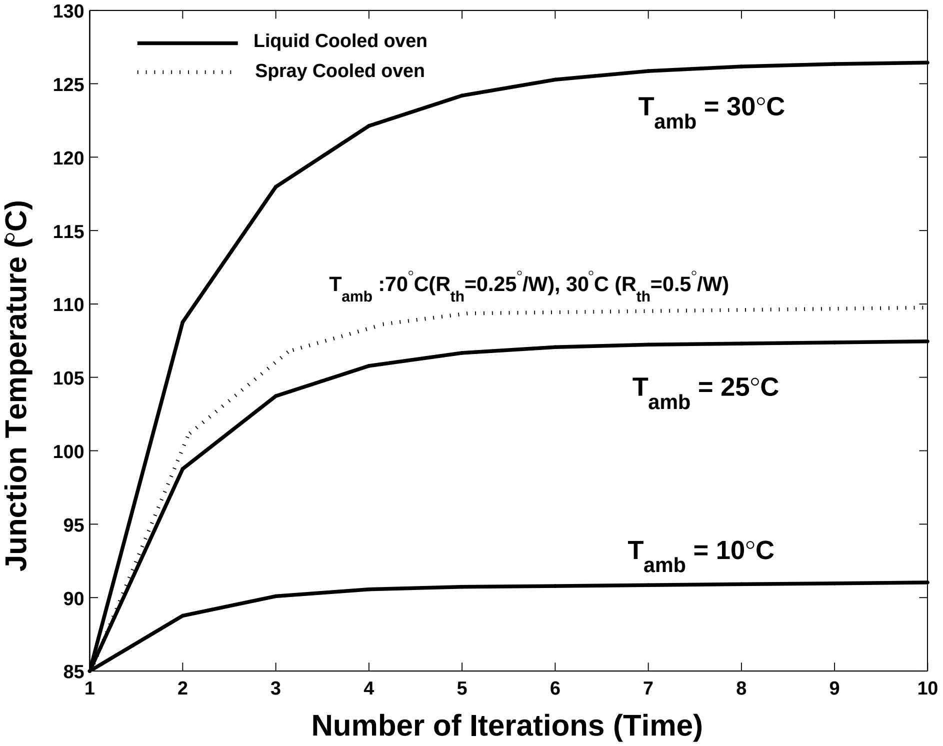


Figure 4.5: Steady-state junction temperature under burn-in conditions for a 130-nm microprocessor with a 10% smaller channel length.

have 10% smaller channel length than the nominal value in 130 nm technology. The excessive leakage due to the smaller channel length increases the junction temperature. As can be seen in Figure 4.5, the ambient temperature must be reduced from 76°C (shown in Figure 4.4) to 30°C (shown in Figure 4.5) for a liquid cooled burn-in oven ($R_{ja} = 0.5°C/W$) to maintain the junction temperature at 110°C. Since it is difficult to maintain the ambient temperature in burn-in ovens around room temperature, it is necessary to reduce the junction to ambient thermal resistance of the burn-in oven. The next generation burn-in ovens are expected to have a thermal resistance of 0.3°C/W using refrigeration as a cooling solution and a thermal resistance of 0.25°C/W using spray cooling technique as cooling solution, respectively. With a thermal resistance of 0.25°C/W, this processor can be burnt-in in the ambient temperature of 70°C (Figure 4.5).

4.3 Thermal Runaway

As mentioned before, temperature and leakage current are strongly correlated and create a positive feedback mechanism between them. Increasing the junction temperature will increase the leakage current and increased leakage current will further increase the junction temperature. Under burn-in or normal operating conditions, designers try to control the junction temperature by removing the heat from the chip. As long as the rate of heat removal is greater or equal to the rate of heat generation, the junction temperature remains constant at the designed operating point. When the rate of heat generation becomes greater than the rate of heat removal, junction temperature starts to increase and thermal runaway occurs. Figure 4.6 shows the transient behavior of the junction to ambient thermal resistance. When the chip is powered on, the thermal resistance starts to increase and reaches to its steady state condition, which as a typical example is 0.6°C/W.

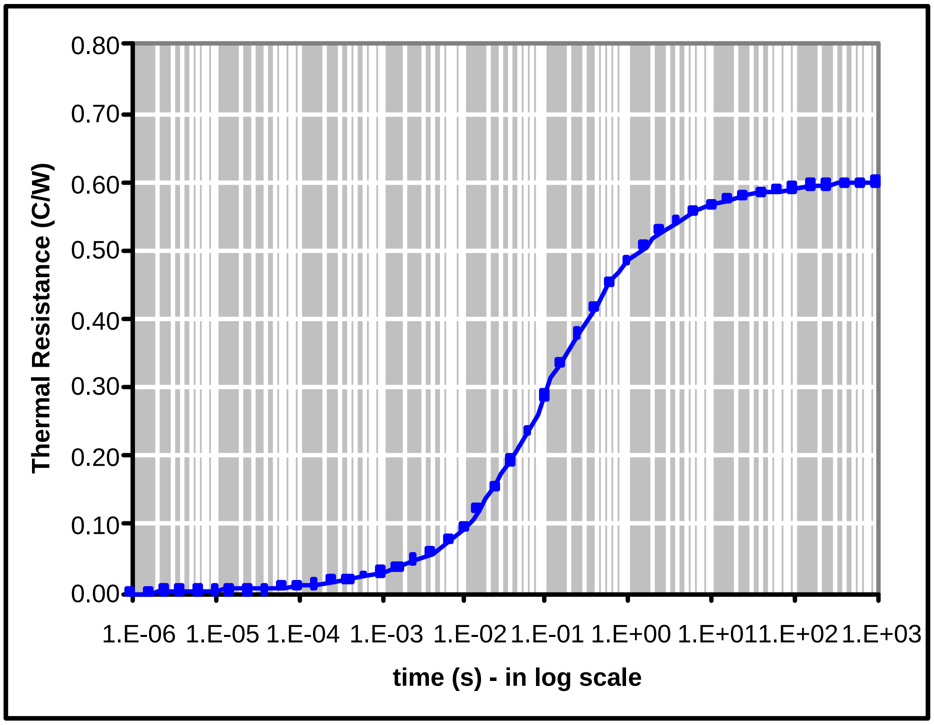


Figure 4.6: Junction to ambient thermal resistance (R_{ja}) increases with time and reaches its steady-state value of $0.6°C/W$.

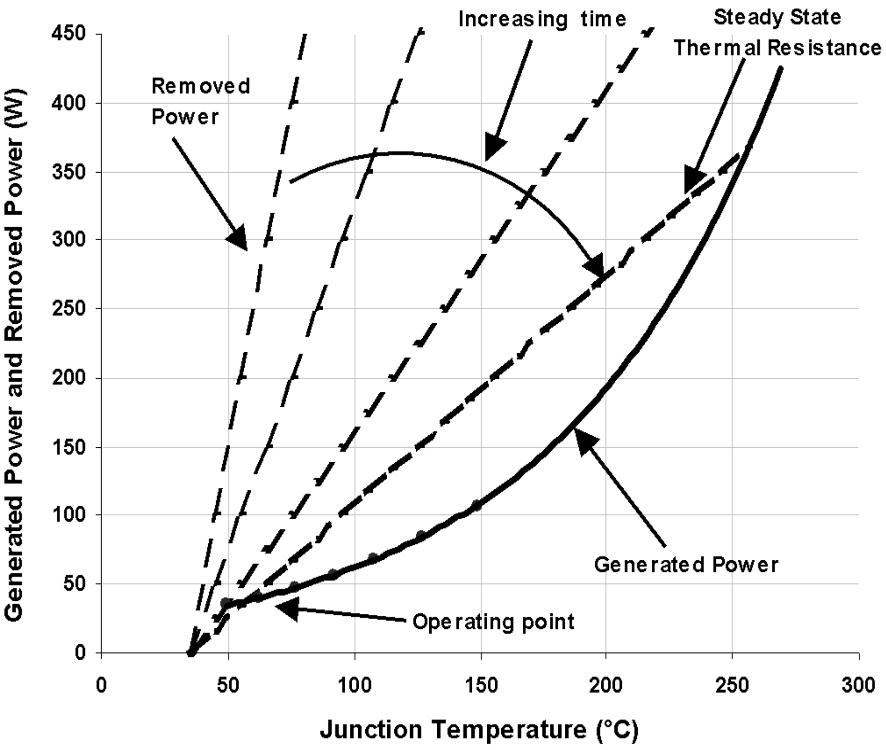


Figure 4.7: 130-nm microprocessor leakage power (exponential) and removed power (straight lines) for a thermal resistance of $0.6°C/W$.

In Figure 4.7, straight lines are drawn using Equation 4.6, which can be expressed as:

$$P_{removed} = \frac{T_j - T_a}{R_{ja}} \tag{4.6}$$

with an ambient temperature of $35°C$. As time increases, the thermal resistance increases from $0°C/W$ and reaches a steady-state value of $0.6°C/W$. Hence, the straight lines represent transient behavior caused by changing thermal resistance with time. On the other hand, the exponential curve is the generated leakage power or chip leakage power at a given ambient temperature. An intersection of the straight line (representing the removed power) and the exponential curve (representing the leakage power) represents the steady state operating condition of the system where removed heat is equal to the generated heat. As long as there is an intersection between the removed power curve and the chip power curve, thermal runaway will not occur.

In Figure 4.8, the leakage power of the chip with nominal leakage and the leakage power of the chip with high leakage due to 10% shorter channel length, versus the junction temperature are depicted. It can be seen that the leakage power for the nominal leakage chip has an intersection with the removed power curve at $110°C$. The slope of the line is $1/0.5°C/W$ and the ambient temperature is $80°C$. At a higher temperature than $110°C$ the removed power is larger than the chip leakage power and at a lower temperature than $110°C$ the leakage power is higher than removed power. This means that from any point in the neighborhood of $110°C$ the temperature will return to $110°C$, which is the design point for burn-in condition. On the other hand if we look at the curve for the high leakage chip, we see that there is no intersection between this curve and the removed power curve with thermal resistance of $0.5°C/W$. Since at all temperatures the removed power is less than the leakage power, for this particular chip, the burn-in environment will lead the chip to thermal runaway. To overcome the problem the burn-in environment must be changed. The new environment is shown in the figure with a thermal resistance of $0.25°C/W$ and an

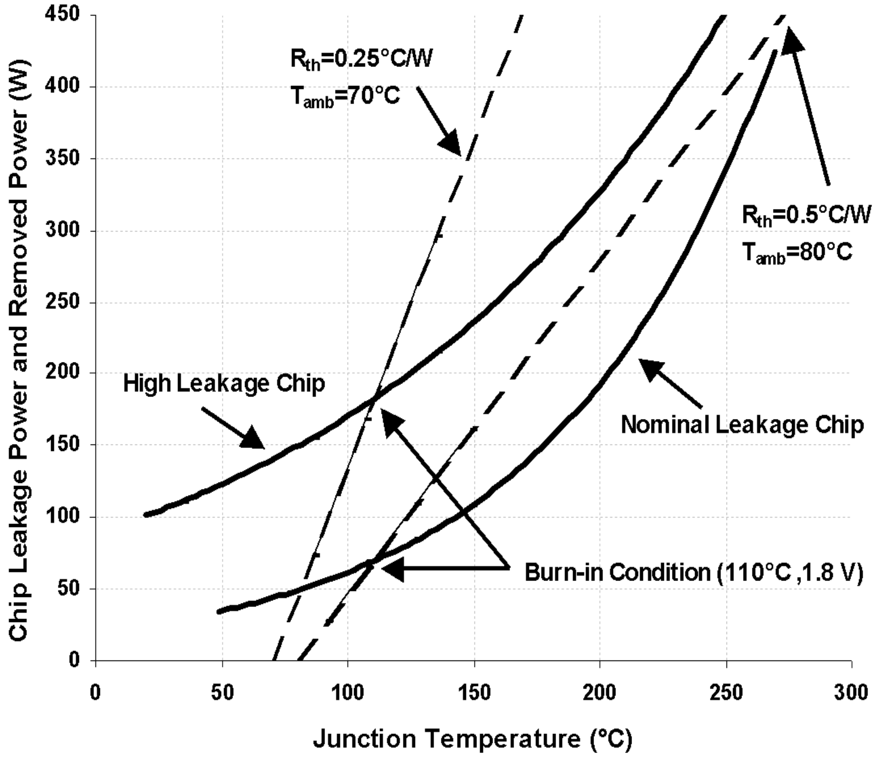


Figure 4.8: Burn-in setup points for nominal leakage and high leakage chips.

ambient temperature of $70°C$. From this experiment, it can be concluded that for scaled chips with higher leakage power, the setup for the burn-in environment must evolve by either reducing either the ambient temperature or the thermal resistance or a combination of both of them. This will shift the removed power curve to the left to intersect the leakage curve of the generated power for the IC at the designed burn-in condition.

4.4 Conclusion

In this chapter, we investigated the thermal management of high performance chips in the burn-in environment. An electro-thermal analysis tool was developed to analyze thermal runaway possibilities due to self-heating in the burn-in environment.

It was concluded that in order to avoid thermal runaway, the burn-in environment must be set up such that the chip power at any temperature higher than the burn-in temperature is less than the removed power so the junction temperature converges to the burn-in design point temperature.

Chapter 5

Microprocessor Design: Optimizations for Low Temperature Operation

Potential advantages of using refrigeration for cooling processors have been reported in the past [4]. Low temperature operation can reduce important device scaling and circuit performance barriers in sub-130nm CMOS technologies. It permits the scaling of supply voltages of high speed circuits to sub-1V by reducing the sub-threshold currents and increasing the carrier mobility in the channels, lowering interconnection resistances significantly, and reducing interconnection related failure machanisms. In this chapter, tradeoffs in microprocessor clock frequency, energy efficiency (MIPS/Watt), die area and system power are investigated when active cooling is used to reduce the operating junction temperature of the microprocessors below a typical hot temperature of 90°C. However, the study is not looking at sub-ambient operating temperatures. It is of interest to lower microprocessor junction temperature below 110°C depending on the cooling efficiency of the technique we have selected.

The purpose of this work is to find out if low temperature CMOS operation has any merit for scaled technologies where transistor subthreshold leakage is relatively high. And if yes, what kind of device, circuit, and design choices are applicable for high performance microprocessors. Consequently, the above mentioned tradeoffs were studied by combining active cooling with:

- Supply voltage (V_{DD}) selection.
- Applying body bias.
- Sizing of transistors in critical and non-critical paths on chip.
- Reduction of channel length (L) as a function of different process technology worst case leakage limits.

Active cooling with and without refrigeration was considered. Several active cooling techniques including air cooling, liquid cooling and refrigeration were investigated. Refrigeration is the most effective cooling solution and is considered for junction temperatures not much below the ambient temperature. Cooling power was considered as part of total system power tradeoffs. System power is the total of chip power (switching and leakage) and power consumed by the cooling system. Analytical models are used for frequency, power, die area, etc. in an electro-thermal analysis tool. The new tool analyzes:

- Frequency, limited by logic and interconnect RC paths.
- System energy efficiency.
- Chip switching and leakage powers, including subthreshold and gate oxide leakage.
- Package and cooling system characteristics.

- Die area.
- Gate oxide reliability-limited maximum V_{cc} constraints.
- Maximum temperature in a self-consistent manner.

The model parameters and input parameters to the tool are typical values. The parameters are extracted from device measurements, process files, and chip measurements.

5.1 Self-Consistent Electro-thermal Optimization

To account for the change in junction temperature, an electro-thermal analysis tool was developed to self-consistently compute the junction temperature, power, and operating frequency of the microprocessor. Figure 5.1 shows the framework of this electro-thermal optimization tool [8]. Starting with an initial assumption of junction temperature ($T_{j-initial}$), the electro-thermal analysis tool first computes frequency and power. At this point, the data from simulation and measurement for I_{on} and I_{off} at the given V_{DD} are incorporated into power and frequency calculations. Other parameters are extracted from process files. In the next step, based on the package and cooling system characteristics, the new junction temperature (T_j) is computed, and the new T_j is the starting point for the power and frequency calculations in the following iterations. These iterations continue until junction temperatures computed in consecutive steps converge to a steady-state temperature. If we do not achieve convergence, it indicates the thermal runaway [84]. When the iterations converge, we obtain the final self-consistent junction temperature. The tool also produces values for corresponding frequency, chip switching and leakage power, active cooling system power, and die area that are consistent with the final temperature. The tool has the following parameters:

- Thermal resistance of the packaging and cooling system.

- Coefficient of performance (COP) for active cooling (defined as the ratio of the cooling power to the power consumed by the cooling system).

- Chip design and process technology characteristics (Figure 5.2).

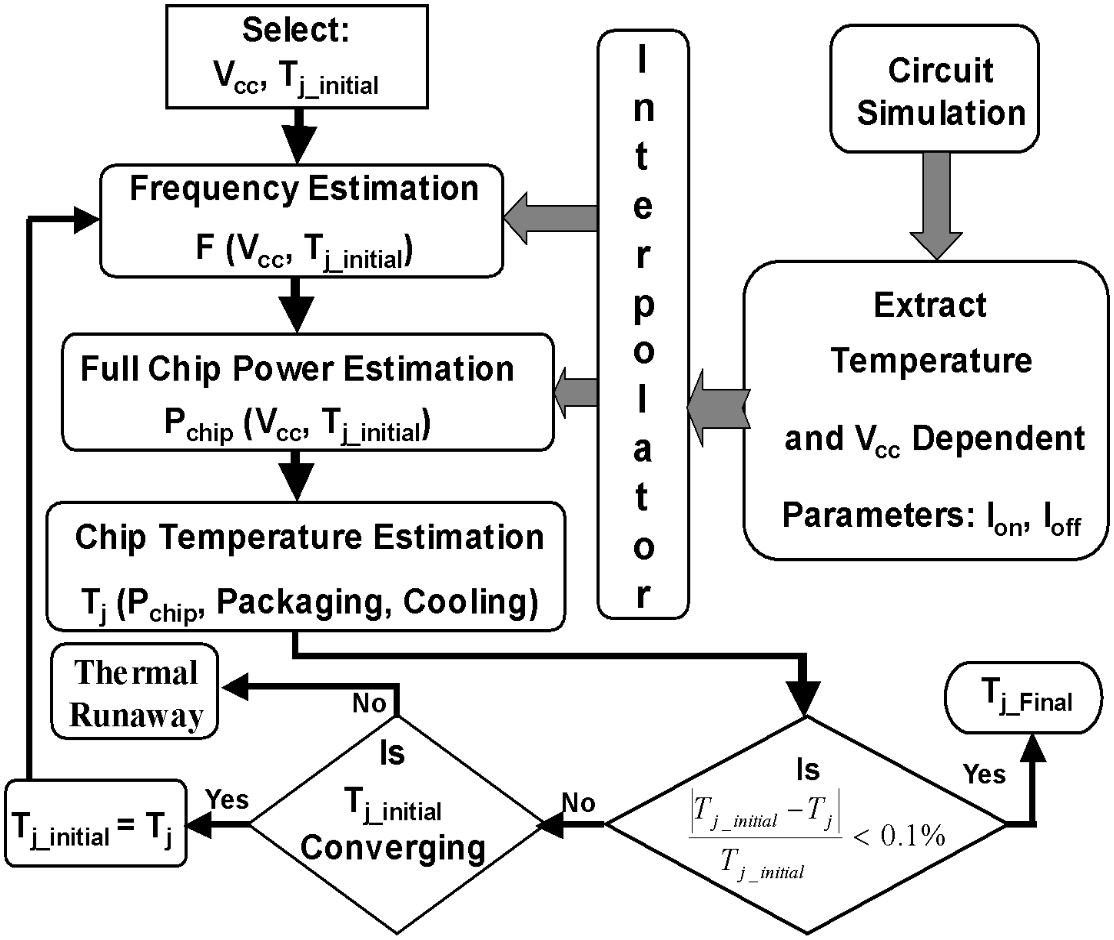

Select:
V_{cc}, $T_{j\_initial}$
Frequency Estimation
F (V_{cc}, $T_{j\_initial}$)
Full Chip Power Estimation
P_{chip} (V_{cc}, $T_{j\_initial}$)
Chip Temperature Estimation
T_j (P_{chip}, Packaging, Cooling)
Interpolator
Circuit Simulation
Extract Temperature and V_{cc} Dependent Parameters: I_{on}, I_{off}
Thermal Runaway
No
Is $T_{j\_initial}$ Converging
Yes
$T_{j\_initial} = T_j$
No
Is $\frac{|T_{j\_initial} - T_j|}{T_{j\_initial}} < 0.1\%$
Yes
T_{j_Final}

Figure 5.1: Framework of the electro-thermal analysis tool for this study [8].

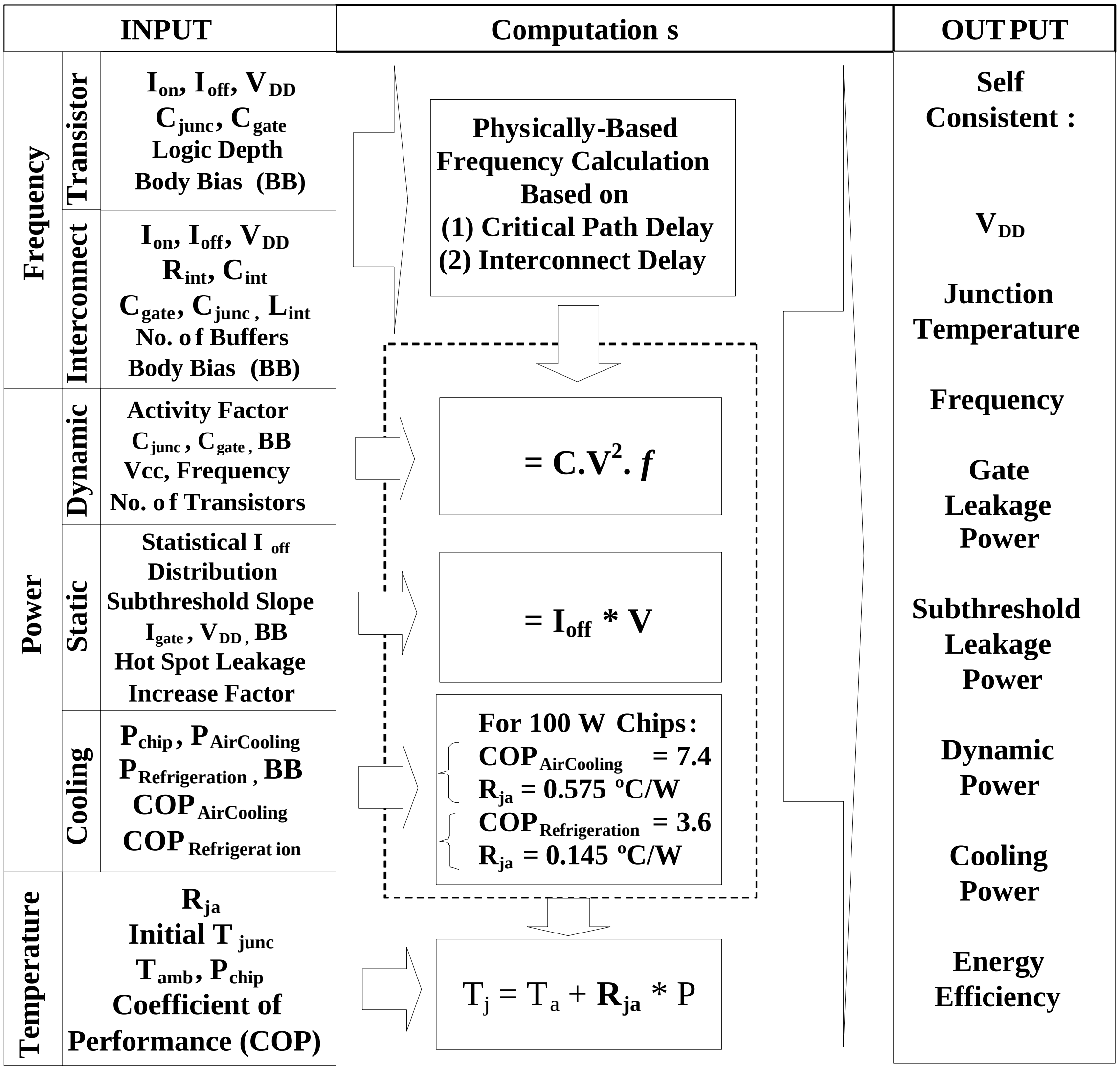


Figure 5.2: Algorithm of the self-consistent physically-based electro-thermal modeling approach.

Although the framwork of this algorithm was well known, we conducted an extensive research to follow each step in this algorithm to keep the tool physically-based. The frequency calculations mimic microprocessor frequency limitations by considering critical path delay and the role of interconnect delay. Circuit parameters such as supply voltage, body bias voltage, number of buffers used in long interconnect lines, and logic depth in critical paths beside transistor parameters including I_{on}, I_{off}, C_j, and C_{gate} and interconnect parameters including C_{int} and R_{int} are extracted, measured, or calculated. The critical path logic depth is used to transition from the transistor to the microprocessor frequency calculation. Figure 5.2 illustrates the important physical parameters which are used to develop the electro-thermal analysis tool. In this figure, in each section (frequency, power and temperature calculations) all the relevant parameters are shown. Each of these parameters, their changes with respect to other parameters, and their impact on the simulation results were studied separately. Their value were extracted from process files, measurements, or physically based calculations. After incorporating these parameters into the tool, the tool was calibrated to actual microprocessor measurements. Figure 5.3 shows the sort level measured data and simulated data for $V_{DD} = 1.1V$ to $V_{DD} = 1.7V$ and $T_j = 25°C$. The calibration was performed by modifing the critical parameters in architectur and circuit level like number of stages in critical path and the activity factor of the chip. Other measurements like total leakage power, total dynamic power, and cooling power were used to confirm the simulation results with respect to measurement results.

5.2 Full Chip Power Estimation

For a 32-bit microprocessor used in this study, the total power can be expressed as:

$$P_{total} = P_{logic} + P_{memory} + P_{I/O} \tag{5.1}$$

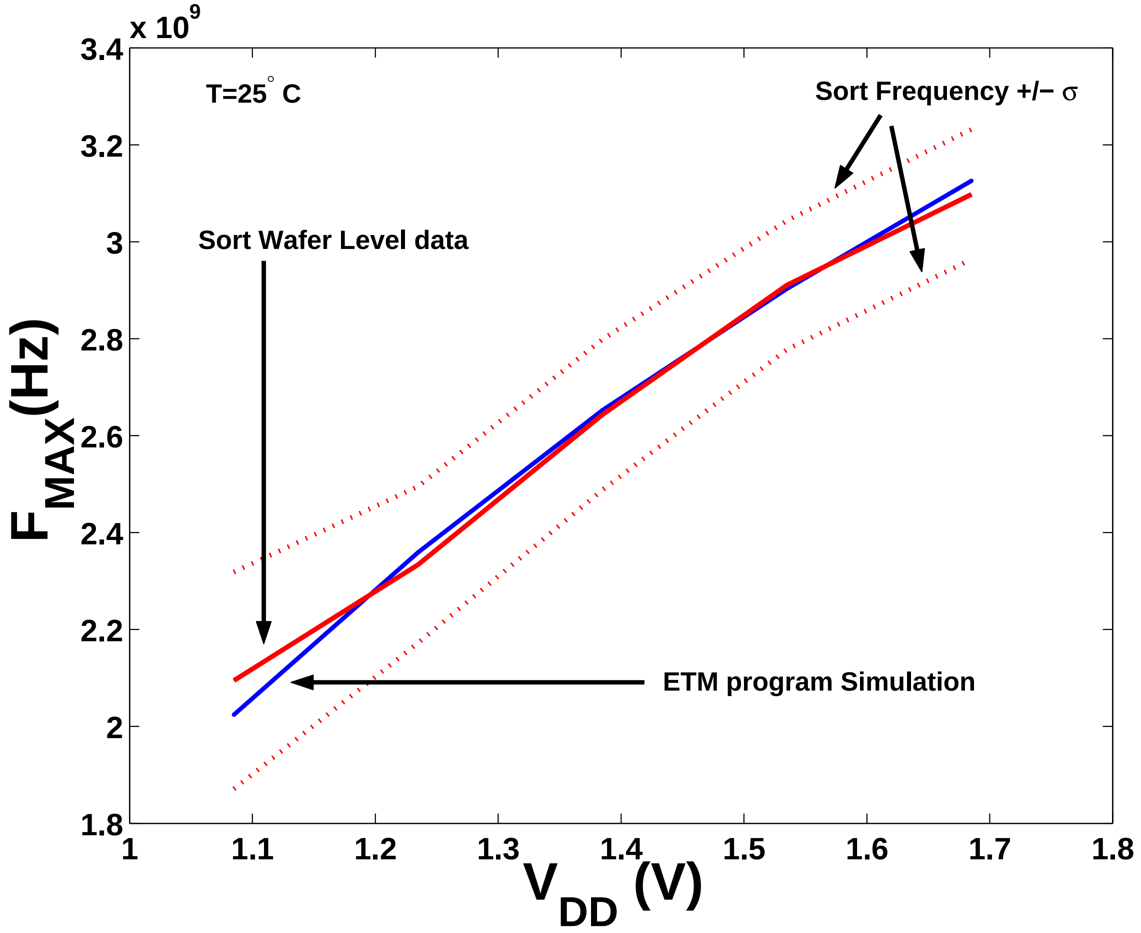


Figure 5.3: Sort level data and measured data for calibration of the electro-thermal tool.

The logic in the microprocessor can be sub-divided into various functional blocks such as the data path, register files and etc. The total width of devices in each block is extracted from the actual design as well as memory (cache) and I/O blocks. For each of these blocks, the leakage power and dynamic power based on the maximum achievable frequency were calculated.

For power calculations, switching power, leakage power and cooling power were incorporated. The short circuit power was ignored. Dynamic switching power was computed according to the appropriate microprocessor activity factor, chip supply voltage, chip switching capacitance, body bias, and the area based on the number of transistors. Static leakage power has also considered gate leakage. The chip leakage is derived based on statistical transistor leakage distributions [59]. The role of hot spots on the chips was also considered for leakage and maximum operating frequency of the chip by giving weight to different blocks on the chip based on the temperature distribution which has been derived from experimental data. Cooling power was computed based on chip power and the COP and thermal resistance of different cooling solutions.

5.2.1 Dynamic Power and Frequency Estimation

The dynamic power or switching power of a general block can be expressed as [66]:

$$\sum_{i=1}^{N} P_{i-switching} = \sum_{i=1}^{N} \alpha_i . C_i . V_{DD}^2 . f(V_{DD}, T_j) \tag{5.2}$$

where N is the number of blocks and α_i is the activity factor and C_i is the total capacitance of the respective block. V_{DD} is the specified operating voltage, f is the the chip operating frequency and T_j is the chip junction temperature. The frequency f can be calculated based on total current charging and discharging the load capacitance as shown in Figure 5.4 and can be defined as [66]:

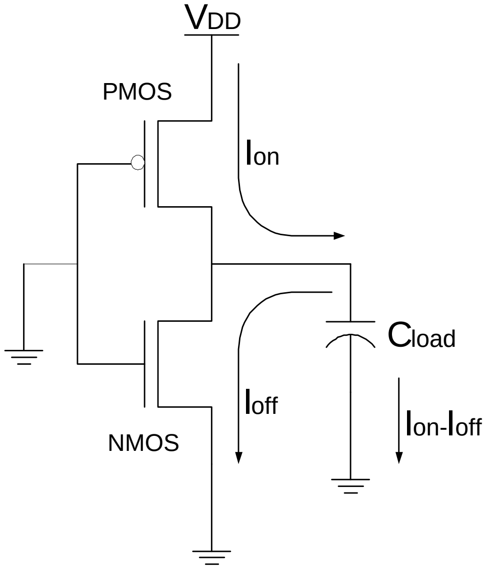


Figure 5.4: The total current that charges the load capacitance in an inverter.

$$\tau = \frac{\tau_{charge} + \tau_{discharge}}{2} \qquad where \qquad \tau_{charge} = \frac{C \times V_{DD}}{I_{charge}} \qquad and \qquad \tau_{discharge} = \frac{C \times V_{DD}}{I_{discharge}}$$

$$\tau = \frac{1}{f} = \frac{(1 + f_{int}).C.V_{DD}}{2.n}(\frac{(I_{on-N}.W_N - I_{off-P}.W_P) + (I_{on-P}.W_P - I_{off-N}.W_N)}{(I_{on-N}.W_N - I_{off-P}.W_P).(I_{on-P}.W_P - I_{off-N}.W_N)} \qquad (5.3)$$

In Equation 5.3, n is the logic depth in the critical path and f_{int} is the fraction of the capacitance contributed by interconnect. I_{on-N} and I_{on-P} are the on-currents of NMOS and PMOS transistors, respectively, and where I_{off-N} and I_{off-P} are the off-currents of these transistors. W_N and W_P are the total widths of NMOS and PMOS devices respectively. Note that the on-current and off-current of transistors are obtained from circuit simulations. In Figure 5.4, C_{load} is charging when the PMOS transistor is ON and NMOS transistor is OFF, so the total current charging the load capacitance will be $(I_{on-P} - I_{off-N})$. The same concept applies for discharging the load capacitance where total current will be $(I_{on-N} - I_{off-P})$.

5.2.2 Capacitance Calculation for Frequency and Power Estimation

The total capacitance is the sum of the two parallel capacitances. One is the capacitance of the driver which drives the load and the other is the capacitance of the load [66].

$$C = C_{load} + C_{driver} \tag{5.4}$$

The driver and load capacitances can be expressed as [66]:

$$C_{driver} = (C_{kja} + C_{kjp} + C_{kjpg}).W_N + C_{ovw-N}.W_N + (C_{kja} + C_{kjp} + C_{kjpj}).W_P + C_{ovw-P}.W_P \tag{5.5}$$

$$C_{load} = fanout.(3Covw + C_{ox}).W_N + fanout.(3Covw + C_{ox}).W_P \tag{5.6}$$

In Equations 5.5 and 5.6 different components can be described as follows:

- C_{kja}: Diffusion to substrate capacitance.
- C_{kjp}: Sidewall capacitance.
- C_{kjpg}: Capacitance between Source/Drain diffusions.
- C_{ox}: The gate (oxide) capacitance.
- C_{ovw}: The overlap capacitance between the gate and the Source/Drain.

The factor of 3 before C_{ovw} in Equation 5.6 accounts for the effective Miller capacitance between the gate and the drain of the NMOS and PMOS transistors, when the load inverter input is high and the output is low or vice versa. Figure 5.5 shows these capacitances in a MOS transistor structure.

For any given block i, the capacitance can be expressed as:

$$C_i = C_{dN}.W_N + C_{dP}.W_P + C_{int} \tag{5.7}$$

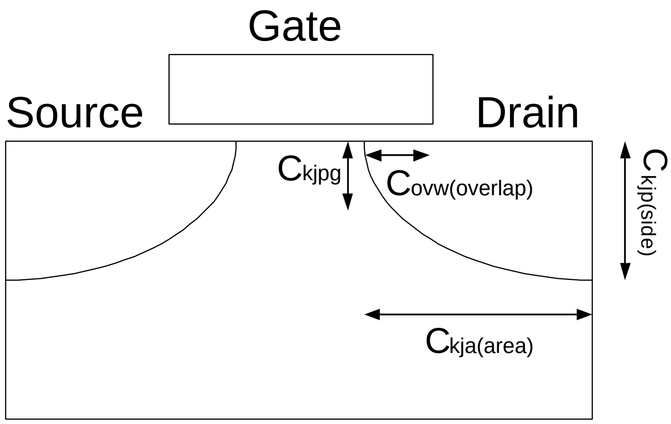


Figure 5.5: Capacitances associated with a MOS transistor.

where:

$$C_{int} = f_1.(C_{dN}.W_N + C_{dP}.W_P) \tag{5.8}$$

and

$$C_d = \beta.(C_{kja} + C_{kjp} + C_{kjpg}) + 2C_{ovw} + C_{ox} \tag{5.9}$$

where β is fitting parameter, $1 \leq \beta \leq 2$. In Equation 5.9, C_{ox} is the gate capacitance.

5.2.3 Leakage Power

Although leakage power can be attributed to both subthreshold and gate leakage, the primarily focus will be on the subthreshold leakage. This is due to the fact that the gate leakage is highly process dependent and can also be tuned to a desirable level by suitable process adjustment. In this work we found that for this technology the gate leakage power was a small percentage of the subthreshold leakage power (less than 1%).

It must be noted that subthreshold leakage is strongly affected by process variations. Process variations cause a variability in the transistor channel length, which in turn causes a variation in the transistor threshold voltage (V_T) due mainly to short-channel effects. The variation in the value of V_T causes a variation in the subthreshold leakage, which can be expressed as [77]:

$$\Delta I_{off} = I_{off-ref}.10^{(V_{T-ref}-V_T)/S} \tag{5.10}$$

In Equation 5.10 V_{T-ref} and $I_{off-ref}$ are the threshold voltage and leakage current respectively at some reference technology node, and S is the subthreshold swing. This indicates that, for a given temperature, as the threshold voltage decreases, the leakage current increases exponentially. To account for the process dependency, the leakage power can be expressed as [59]:

$$\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{I_{off}^{3\sigma}(V_{DD}, T_j)}{m}.exp[\frac{\sigma^2}{2\lambda^2(V_{DD}, T_j)} - \frac{3\sigma}{\lambda(V_{DD}, T_j)}].W_i.X_n.V_{DD} \tag{5.11}$$

where

- n is the number of blocks.
- m is fraction of off devices.
- X_n is the noise factor I_{off}.
- σ is the standard deviation of I_{off} process distribution.
- λ is the slope of the I_{off} vs. channel length (L) curve.
- W_i is the total width of transistors in block i.

5.3 Reliability and Cooling Constraints

This section will describe that how the gate oxide reliability and cooling constraints are integrated into the electro-thermal tool.

5.3.1 Reliability Constraints

In the electro-thermal modeling of the chip, it is necessary to consider the long term gate oxide reliability of the chip. In any given junction temperature there is a maximum V_{DD}, beyond which, the gate oxide reliability of the chip will be compromised. To validate the V_{DD} values in the self-consistent methodology, a gate oxide reliability constraint equation is used as given below:

$$V_{DD} \leq V_{max} = T_j \times R + c \qquad and \qquad R = \frac{xmV}{1°C} \tag{5.12}$$

In Equation 5.12, R is a technology dependent reliability factor. The constant c can be calculated based on R and nominal values of V_{DD} and T_j. V_{max} is the maximum voltage for V_{DD} that satisfies the reliability criterion for gate oxide. The reliability criterion is checked at the end of each iterative loop in Figure 5.1 in calculating T_j and V_{DD}.

5.3.2 Cooling Constraints

The operating junction temperature of a chip depends on the cooling solution that is used for conducting the generated heat from the junction to the ambient surrounding the chip. Different cooling solution can be used to remove the generated heat. In this study we focus on the air cooling and refrigeration to compare the low temperature and high temperature operation of microprocessors. The model that is used in the electro-thermal tool is shown in Figure 5.6. In this model:

$$T_j - T_{amb} = P_{chip}.R_{ja} \qquad where \qquad R_{ja} = R_{jc} + R_{ca}$$

$$T_{amb} - T_{out} = P_{sys}.R_{sys}$$

$$P_{cooling} = Q_{elec} = \frac{P_{chip}}{COP} \qquad and \qquad \eta = 1 - \frac{1}{COP}$$

$$P_{sys} = P_{chip} + P_{cooling} = P_{chip} + (1 - \eta).P_{chip} = (2 - \eta).P_{chip} \tag{5.13}$$

In Equation 5.13, P_{sys} is the total system power which includes total chip power (dynamic and static) as well as power spent due to any dynamic cooling mechanism with an efficiency of η ($\eta < 1$). T_{amb} is the ambient temperature (temperature immediately outside the chip case) and T_{out} is the external room temperature. Table 5.1 shows the cooling parameters for different cooling techniques for a 100 W processor. In this table R_{j-c} is the junction to case and R_{c-a} is the case to ambient thermal resistance. Q_{elec} is the amount of power used by the cooling system and the coefficient of performance (COP) is the ratio of the chip power to cooling power. T_a and T_j are ambient and junction temperature respectively.

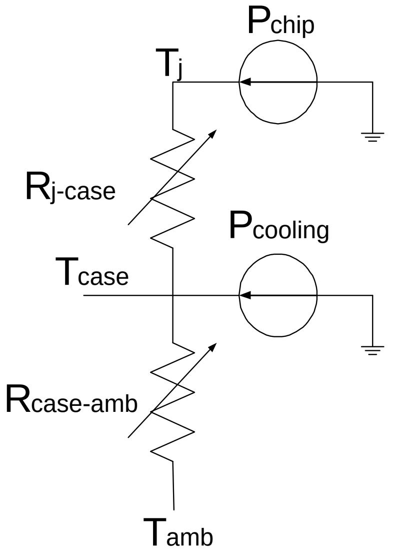


Figure 5.6: Thermal circuit illustrating the relationship between the chip and system level power, thermal impedance and temperature.

	Air Cooling	**Liquid Cooling**	**Refrigeration**
R_{j-c} $(°C/W)$	0.305	0.305	0.305
R_{c-a} $(°C/W)$	0.393	0.290	0.050
Q_{elec}	2.0	7.5	50.0
COP	50.0	13.3	2.0
T_a $(°C)$	35.0	35.0	35.0
T_j $(°C)$	104.8	94.5	70.5

Table 5.1: Cooling parameters for a 100 W processor which is cooled in 3 different ways: air cooling, liquid cooling and refrigeration.

5.4 Optimization of Microprocessor Operating Frequency Subject to Reliability Constraints

To demonstrate how the modeling works, Figure 5.7 shows the results of an optimization for an example microprocessor in a low-leakage 130nm process technology for a typical package and air cooling system. Solutions are obtained for different V_{DD} values, and an operating point is accepted only if V_{DD} does not exceed the gate oxide reliability-limited maximum (V_{max}) at the final T_j. Therefore, reliability considerations set the maximum allowable supply voltage. The highest optimal frequency and corresponding T_j are set by $V_{DD} = V_{max}$ at that T_j. For these simulations, the ambient temperature (T_a) is set to $35°C$.

The X-axis of Figure 5.7 represents the chip operating frequency. The Y-axis has captured multiple parameters including power, temperature, supply voltage and the reliability maximum allowed supply voltage (V_{max}). As the supply voltage increases, the

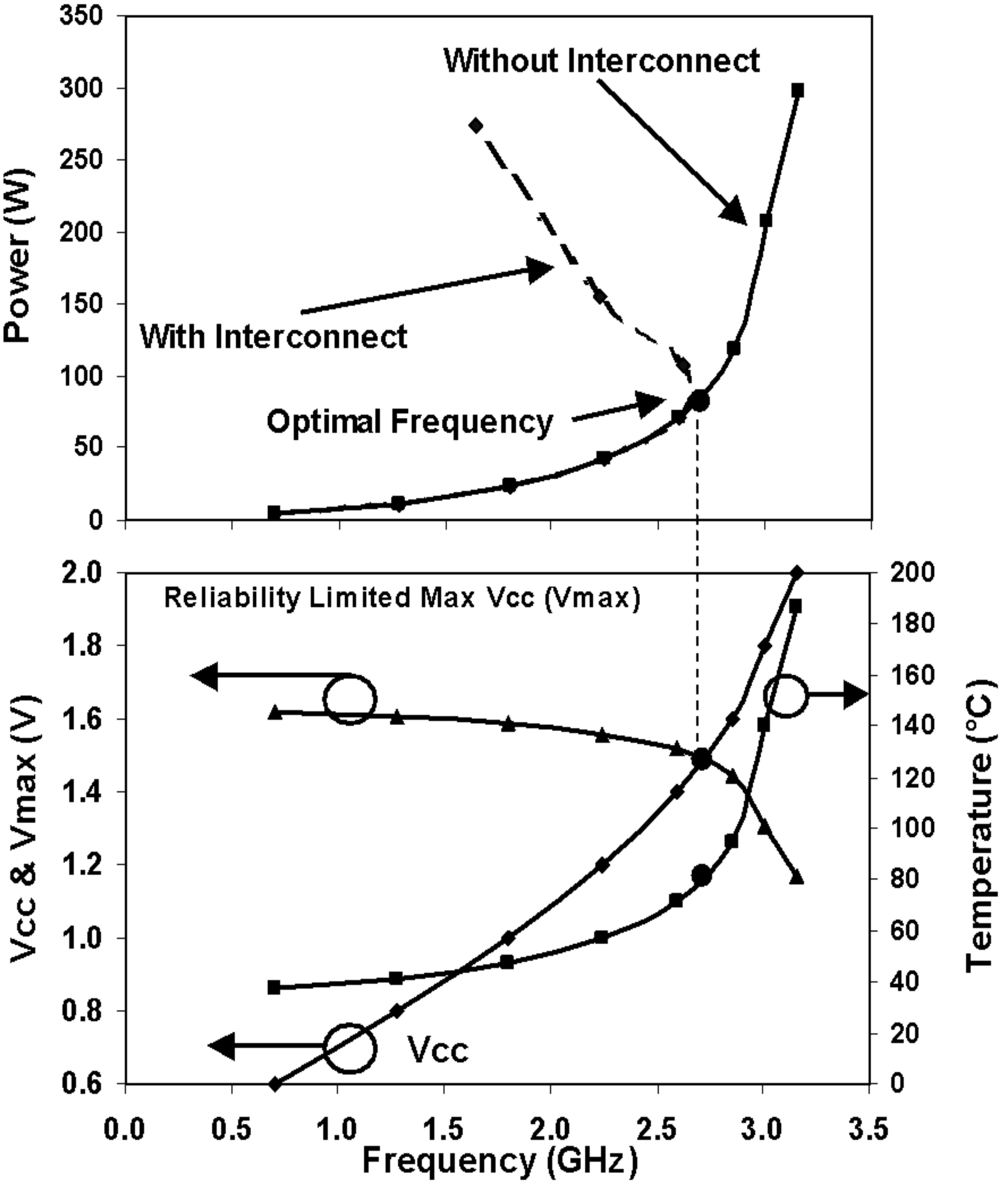


Figure 5.7: Optimization of microprocessor operating frequency subject to reliability constraints.

chip frequency increases, and the junction temperature rises. However, the maximum reliability-limited supply voltage reduces at higher temperatures due to degraded gate oxide reliability performance. Consequently the maximum supply voltage is determined at the intersection of the V_{DD} and reliability limited maximum V_{DD} (V_{max}) curves. This sets the junction temperature, frequency and power of the chip accordingly. The optimal operating frequency is $2.7GHz$ at V_{DD} of 1.5V and T_j of $81°C$ where the system power is $82W$. Interconnect RC delays with repeaters can also limit the maximum frequency (top portion of Figure 5.7) since RC delays change with T_j in a different way from transistor performance and circuit delay. Also, RC delay is relatively insensitive to V_{DD} change, whereas circuit delays in logic paths change significantly with V_{DD}. In an optimum design interconnect should not limit chip frequency and power at the optimum frequency. This allows transistors to provide their highest potential performance. This is shown in top portion of Figure 5.7 where interconnect dashed line increases power without improving the chip frequency after optimal operating point.

5.5 Low-leakage vs. High-leakage Technology Trade-offs

Figure 5.8 shows the frequency and power tradeoffs for iso-reliability high performance operation and iso-power operation conditions when refrigeration active cooling is incorporated. The relative contributions of cooling power, dynamic power, and leakage power demonstrate how leakage power and cooling power can be traded off. This is best shown in the iso-power case. For a constant power limit of 80W, the frequency increases by 4.5% going from air cooling to refrigeration in a low-leakage technology, and by 7.5% for high-leakage technology (Figure 5.8). This happens because when leakage is a large percentage of the total power (31% in this case), the leakage power reduction due to lower T_j translates to more savings in total chip power. Then, power overhead of the cooling system will have less impact on total system power. To achieve the highest operating frequency in line with microprocessor applications, the iso-reliability case must be studied as shown in Figure 5.8. For a low-leakage technology, the reliability-limited frequency improves by 12% and the system power increases by 35% going from air cooling to refrigeration. When leakage is higher, the frequency increases by 17% for a 62% increase in system power. Therefore, the frequency vs. power tradeoff is worse when the leakage is higher. Frequency improvements in both cases come from operation at reduced temperature and the higher V_{DD} allowed at lower T_j due to utilizing refrigeration.

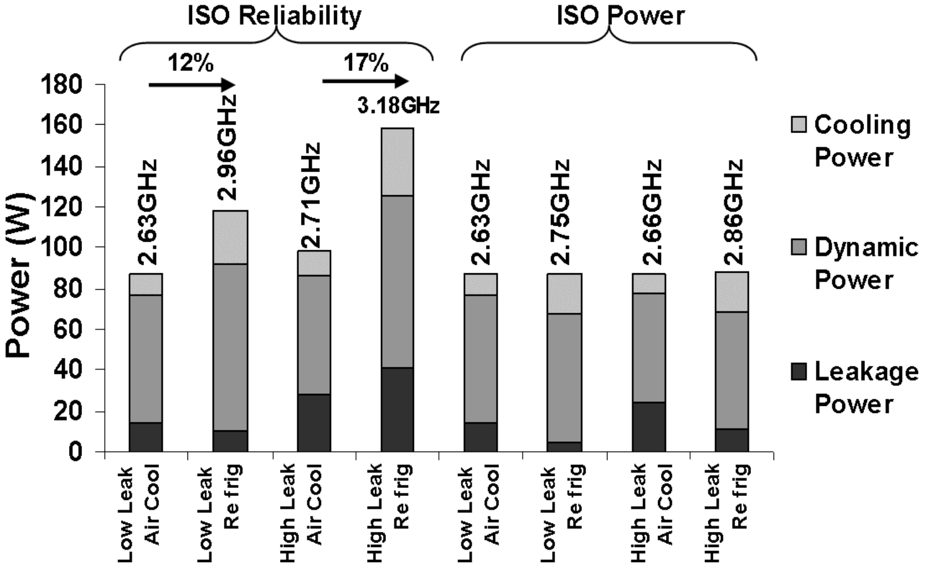


Figure 5.8: Reliability and power limited maximum frequency achievable for low and high leakage technologies with refrigeration.

5.6 Power, Frequency and Energy Comparisons for Optimal Design at Low Temperature

Now that the active cooling for different amounts of worst-case process technology leakage constraints have been studied, the optimum design for low temperature CMOS operation can be investigated. The following design techniques for optimal low temperature operation are considered: Changing V_{DD}, changing transistor channel length and enhancing the process technology, changing transistor sizing, and applying body bias. Figure 5.9 shows tradeoffs in system power, energy efficiency and die area vs. frequency offered by forward body bias (FBB), shortening L, changing V_{DD} and transistor sizing, with and without refrigeration. System power and system energy efficiency as a function of chip frequency is plotted in Figure 5.9. These graphs are normalized to air cooling power, energy efficiency and frequency. When refrigeration combined with a design technique is utilized, the goal is to minimize the slope in the system power versus chip frequency curves. This corresponds to maximum chip frequency increase for lowest increase in system power. For system energy efficiency, the goal is the maximum change in frequency and highest possible energy efficiency.

Figure 5.9 shows how applying forward body bias in addition to refrigeration increases the frequency but the rate of system power increase is rather steep. Applying 0.4V FBB increases frequency by an additional 2.7% and increases power by 27%. The best FBB tradeoff is when its value is limited to 100mV. FBB also degrades energy efficiency by 16%. Decreasing V_{DD} from 1.56V to 1.4V lowers both frequency and system power. However, at lower V_{DD} values, the rate of chip slowdown is much higher than the achieved power saving. Reducing sizing by lowering transistor width has similar tradeoffs as for the supply voltage.

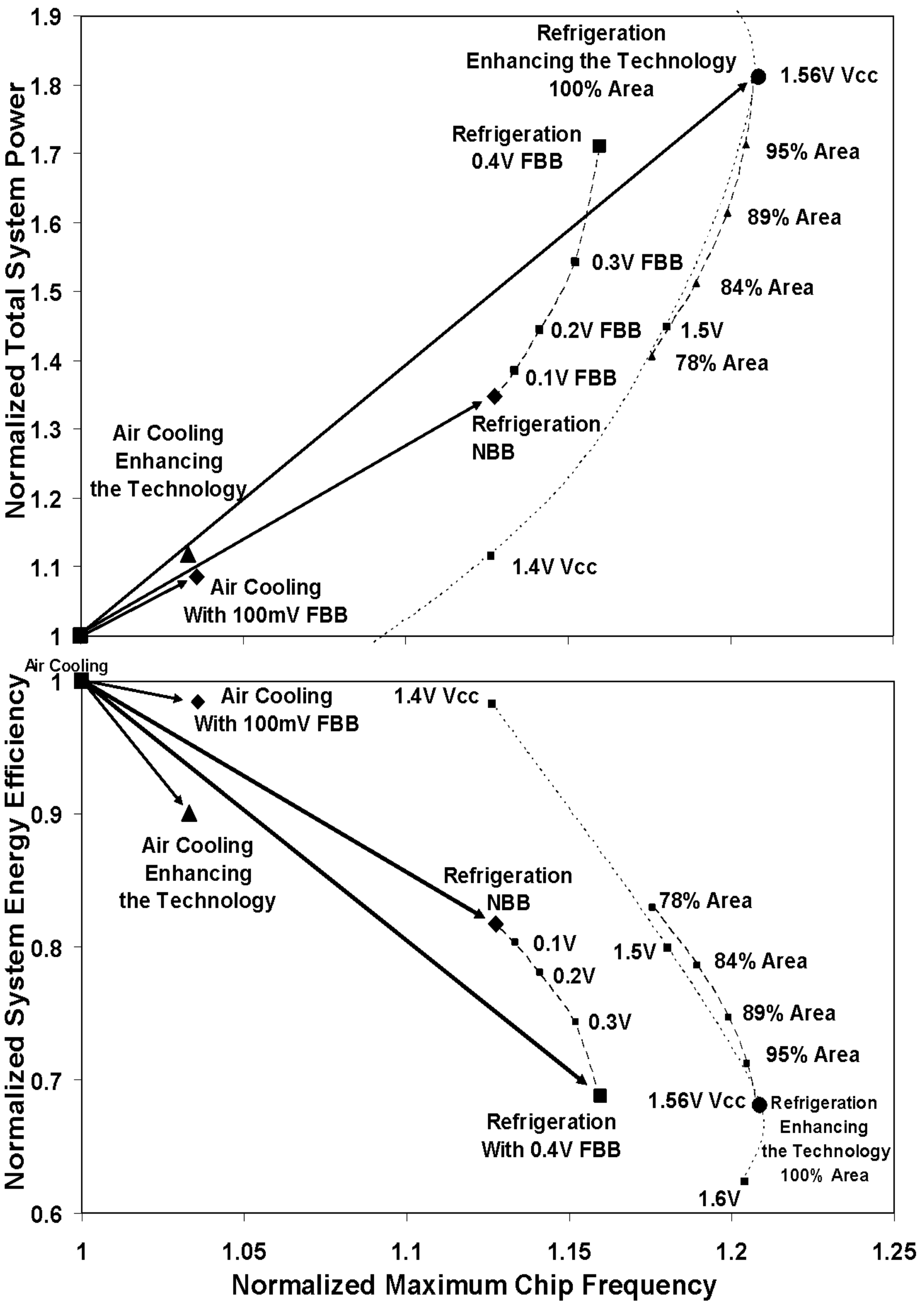


Figure 5.9: Tradeoffs in system power, energy efficiency, die area, and frequency by different circuit and design techniques.

Table 5.2 summarizes integration of different design solutions and explores the design space for iso-power and iso-frequency conditions. Combined refrigeration with shorter L (enhancing technology), appropriate V_{DD} selection and transistor sizing provides the highest frequency for any system power limit and the highest energy efficiency for any target frequency. The greatest frequency increase of 11% is achieved for the iso-power case at a V_{DD} of 1.41V, a temperature of $31°C$ and 11% smaller area for enhancing the technology in our design space. While performing iso-frequency analysis, enhancing the technology (shorter L), provides 38% total system power saving at a V_{DD} of 1.36V, a temperature of $15°C$ and 33% smaller area. In both cases we improve the energy efficiency by 11% and 62%, respectively.

In summary, Table 5.2 shows improvements in frequency for equal power and reduction in power for a specific target frequency for air cooling and refrigeration when transistor sizing and the supply voltage are optimized for optimal forward body bias and shorter L. Die area changes are also compared. Reducing L provides better frequency and power improvement than FBB in all cases.

Also, combining refrigeration with shorter L is the best for minimizing power and maximizing frequency. Furthermore, this strategy provides lowest die area when comparing power at equal frequency, and second best when comparing frequency at equal power.

5.7 Conclusion

In this work an electro-thermal analysis tool was developed. Using this tool the tradeoffs in microprocessor frequency and system power achievable by combining refrigeration with supply voltage selection, body bias, transistor sizing, and shorter channel length were studied. Reducing the channel length provides better frequency and power improvement

Iso Power	V_{DD}	T_j	**Area**	**Leakage**	**Cooling**	**Energy**	**Frequency**
	(V)	$(°C)$				**Efficiency**	
Air Cooling (Reference)	1.49	80	100%	19%	14%	100%	100%
Air Cooling (100mV FBB)	1.47	81	95%	21%	14%	102%	102%
Air Cooling (Enhanced Technology)	1.40	82	89%	31%	14%	107%	107%
Refrigeration (200mV FBB)	1.60	30	73%	13%	29%	108%	108%
Refrigeration (Enhanced Technology)	1.41	31	89%	18%	28%	111%	111%

Iso Frequency	V_{DD}	**Temp**	**Area**	**Leakage**	**Cooling**	**Energy**	**Power**
	(V)	$(°C)$				**Efficiency**	
Air Cooling (Reference)	1.49	80	100%	19%	14%	100%	100%
Air Cooling (100mV FBB)	1.45	76	89%	19%	14%	115%	87%
Refrigeration (200mV FBB)	1.46	20	73%	8%	29%	134%	75%
Air Cooling (Enhanced Technology)	1.34	72	78%	27%	14%	137%	73%
Refrigeration (Enhanced Technology)	1.36	15	67%	12%	30%	162%	62%

Table 5.2: Optimum design space for active cooling at iso-power and iso-frequency conditions in an ambient temperature of $35°C$.

than forward body bias. Also, combining refrigeration with shorter channel length produces the best power-frequency tradeoffs.

Chapter 6

Conclusion

Heat and power management of high performance VLSIs is becoming one of the most important issues for scaled CMOS technologies. These issues involve power and junction temperature estimation for normal and stress conditions, long term reliability in normal operating conditions and reliability screening of the chip under stress conditions. In this thesis, some of these issues were described and some novel methodologies to address them were developed.

In chapter two, burn-in as a reliability screening test and the burn-in issues with respect to technology scaling were discussed. In chapter three, after reviewing the concept of the thermal resistance of the CMOS, a novel technique was introduced to estimate the junction temperature in normal and burn-in conditions. The technique was used for burn-in optimization with respect to reliability and yield. In chapter four a new insight for thermal runaway as a threat to the yield of VLSI chips during burn-in was discussed. Finally in chapter five a self-consistant electro-thermal modeling tool was developed to study the tradeoffs of low temperature operation. This model was described and the result of a low temperature operation study was presented.

6.1 Thesis Contribution

This thesis has made several contributions in the design, quality and reliability of integrated circuits.

Contemporary VLSI designs have become extremely power hungry and, as a consequence, their junction temperature has increased with scaling. The increased junction temperature in scaled technologies and its effects on the reliability, quality and performance of the circuits, have been the primary reason for designers to focus on junction temperature estimation in the early stages of design. In this thesis, a technique for junction temperature estimation is developed. Using this technique, the increase in the normalized junction temperature with scaling was predicted. Based on this particular work, papers J2, J3, and C4 were published. This included an invited paper to IEEE Transaction on Device and Materials Reliability.

Unabated increase in junction temperature is the cause of several quality and reliability problems, including thermal runaway. Therefore, we must devise technology, circuit, and operational techniques to contain this increase. This thesis provide a new insight into the concept of thermal runaway and how it may best be avoided. Based on this work, papers J1, C2, and C3 were published.

One of the possible ways to avoid the static power increase is low temperature operation, which provides lower static power and higher performance. In this thesis an electro-thermal tool was developed to study the low temperature operation of the high performance processors. In this tool all the physical parameters of the chip at device, circuit and system level was incorporated and the tool was calibrated to an actual microprocessor. The result of this work was published in paper C1 as an invited paper in IEEE DAC. The complete list of the publications is presented in appendix A.

6.2 Future Work

Power dissipation limits have emerged as a major constraint in the design of high performance circuits such as processors. At the low end of the performance spectrum, namely in the world of handheld and portable devices or systems, power has always dominated over performance (execution time) as the primary design issue. Battery life and system cost constraints drive the design team to consider power over performance in such a scenario. On the other end of the spectrum, the total power consumption of high performance microprocessors increases with scaling. Off-stage leakage current is an increasing percentage of the total current at the 130 nm and sub-100 nm nodes under nominal conditions. For 130 nm technology the leakage power is 20% to 50% of the total power and is expected to increase to even more than 50% for sub-100 nm technologies.

Moreover in a reliability screening environment (e.g. burn-in) where ICs are tested under voltage and temperature stress, the ratio of leakage to active power becomes adverse and increases the probability of thermal runaway. These issues must be addressed at the architectural, circuit design and packaging levels. Such a scenario is not just limited to high-end processors. High performance analog and mixed signal circuits, and memories are also confronted with similar challenges. For example, multi GHz clock and data recovery circuits, line drivers, back plane drivers dissipate significant amount of power. Similarly, Content Addressable Memories (CAM) due to their parallel search capabilities also dissipate large amount of power. In other words, thermal management in high performance VLSI circuits will become an integral part of the design, test, and manufacturing.

The followings areas need extensive research to address the above mentioned issues. The first step to address these issues is to implement more powerful CAD tools to estimate the total power consumption and junction temperature of the ICs under nominal and stress conditions. These tools must incorporate the circuit, architectural, and packaging characteristics of the IC to study power, energy efficiency, and performance tradeoffs. I will continue my research in this area as an integrated part of my research in other areas.

6.2.1 Low Power and High Performance VLSI Design Using Circuit and Layout Techniques

The reliability of circuits is exponentially dependent on the operating temperature. Even small differences in operating temperature (of the order of $10 - 15°C$) can result in a factor of 2 reduction in the device lifetime. Besides reliability, thermal analysis is also important because cross-chip temperature gradients and thermal coupling effects induced by localized power dissipation may affect the performance of the circuit. Performance degradation caused by thermally-induced device mismatch is a major concern in the design of integrated circuits, particularly in circuits which experience a large amount of dissipated power, or in high precision circuits, such as data converters, instrumentation amplifiers, analog multipliers, etc. Given the above, it is not surprising that the awareness of thermal issues and the need for thermal co-design has increased over the past few years.

Part of my future plan will be focusing on the thermal behavior of integrated circuits and improving them effectively by means of layout optimization and/or circuit and technology techniques. Layout and circuit techniques can reduce the junction temperature and leakage power, which are strongly correlated.

6.2.2 Packaging Research for High Performance Systems on Chip

The current advances in packaging technology, especially in Chip Scale Packages (CSP), are leading to packages with smaller pitches and more complex designs. An efficient electro-thermal model must let the designer take into account the heat transfer within devices and surrounding materials to silicon chip and advanced package-level cooling solutions. Incorporating package-level cooling solutions will give designers the capability of studying the low temperature operation of ICs.

Other than air-cooling, liquid cooling and refrigeration seem to be good candidates for low temperature operation. Another part of my research will focus on low temperature operation, where the designers will benefit from higher performance and lower power consumption at the circuit level. However, the cost of these techniques must be evaluated and power (cost)/performance tradeoffs must be studied at the package level rather than circuit level alone.

Appendix A

List of Publications

A.1 Journal Papers:

J1- A. Vassighi, M. Sachdev, ”Thermal Runaway: A new challenge for Deep sub-micron Technologies,” submitted to IEEE Design and Test of Computers.

J2- A. Vassighi, O. Semenov, M. Sachdev, A. Keshavarzi, C.F. Hawkins, ”CMOS IC Technology Scaling and its Impact on Burn-in,” Invited Paper, To appear in IEEE Transactions on Device and Materials Reliability, 2004

J3- O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi and C.F. Hawkins, ”Effect of CMOS technology scaling on thermal management during burn-in,” IEEE Transactions on Semiconductor Manufacturing, Vol. 16, No. 4, pp. 686- 695, Nov. 2003.

J4- O. Semenov, A. Vassighi and M. Sachdev, ”Leakage current in sub-quarter micron MOSFET: A perspective on stressed delta IDDQ testing,” Journal of Electronic Testing (JETTA), Vol. 19, No.3, pp. 341-352, 2003.

J5- O. Semenov, A. Vassighi and M. Sachdev, ”Impact of technology scaling on thermal behavior of leakage current in sub-quarter micron MOSFETs: Perspective of low temperature current testing,” Microelectronics Journal, Vol. 33, No. 11, pp. 985-994, 2002.

A.2 Referred Conference Papers:

C1- A. Vassighi, A. Keshavarzi, S. Narendra, G. Schrom, Y. Ye, S. Lee, G. Chrysler, M. Sachdev, V. De, ”Design optimizations for microprocessors at low temperature,” Invited Paper, Proc. Design Automation Conference (DAC), pp. 2-5, June, 2004.

C2- A. Vassighi, O. Semenov and M. Sachdev, ”Thermal runaway avoidance during burn-in,” IEEE IRPS 2004, Phoenix, pp. 655-656, April 25 - 29, 2004.

C3- A. Vassighi, O. Semenov, M. Sachdev and A. Keshavarzi, ”Thermal management of high performance microprocessors in burn-in environment,” IEEE Int. Symposium on Defect and Fault Tolerance in VLSI Systems (DFT’03), Cambridge, MA, USA, pp. 313-319, Nov. 2003.

C4- O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi and C.F. Hawkins, ”Burn-in temperature projections for deep sub-micron technologies,” IEEE Int. Test Conference, pp. 95-104, Oct. 2003.

C5- A. Vassighi, O. Semenov, M. Sachdev and A. Keshavarzi, ”Effect of static power dissipation in burn-in environment on yield of VLSI,” IEEE Int. Symposium on Defect and Fault Tolerance in VLSI Systems (DFT’02), Vancouver, Canada, pp. 12-19, Nov. 2002.

C6- A. Vassighi, O. Semenov, M. Sachdev and A. Keshavarzi, "Impact of power dissipation on burn-in test environment for sub-micron technologies," IEEE Int. Workshop on Yield Optimization and Test (YOT), pp. 1-5, Oct. 2001.

Appendix B

Glossary of Terms

VLSI Very Large Scale Integrated.

CMOS Complementary Metal Oxide Semiconductor.

TTL Transistor-Transistor Logic.

ESD Electro-Static Discharge.

DFR Decreasing Failure Rate.

CFR Constant Failure Rate.

IFR Increasing Failure Rate.

BI Burn-in.

WLBI Wafer Level Burn-in.

BIB Burn-in Board.

DUT Device Under Test.

TDBI Test During Burn-in.

TDDB Time-Dependent Dielectric Breakdown.

SILC Stress-Induced Leakage Current.

MTTF Mean Time To Failure.

EM Electro-Migration.

DPM Device Per Million.

FIT Failure In Time.

MTCMOS Multi-Threshold CMOS.

RBB Reverse Body Bias.

GIDL Gate Induced Drain Leakage.

DCT Discrete Cosine Transformation.

W Width of CMOS transistor.

L Channel Length of CMOS transistor.

R_{ja} Junction to ambient thermal resistance.

LP Low Power.

HP High Performance.

UHP Ultra High Performance.

RTD Resistive Temperature Detectors.

MIPS Million Instruction Per Second.

COP Coefficient Of Performance.

FBB Forward Body Bias.

CSP Chip Scale Package.

Bibliography

[1] "IBM 6X86MX Microprocessor". http://www-3.ibm.com/chips/techlib/techlib.nsf/techdocs/AF16346AD95E76D987256A310064E3B4.

[2] "Pentium Procesor with MMX Technology". http://cs.mipt.ru/docs/comp/eng/hardware/processors/intel/i586/p55/main.pdf.

[3] "Panel on Burn-in Elimination". *International Reliability Physics Symposium (IRPS)*, 2001.

[4] I. Aller, K. Ghoshal, H. Schettler, S. Schuster, Y. Taur, and D. Torreiter. "CMOS Circuit Technology for Sub-Ambient Temperature Operation". *IEEE International Solid-State Circuits Conference*, pages 214–215, 2000.

[5] A. Alvandpour, R. Krishnamurthy, S. Borkar, A. Rahman, and C. Webb. "A burn-in tolerant dynamic circuit technique". *IEEE Custom Integrated Circuits Conference*, pages 81–84, 2002.

[6] Mark Miller (AMD). "Next generation burn-in and test systems for Athlon microprocessors: hybrid burn-in". *Burn-in and Test Socket Workshop*, 2001.

[7] K. Banerjee and R. Mahajan. *Intel Development Forum*, 2002. ftp://download.intel.com/research/silicon/Thermals-press-IDF-0902.pdf.

[8] K. Banerjee, L. Sheng-Chih, A. Keshavarzi, S. Narendra, and V. De. “A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management”. *IEEE International Electron Devices Meeting*, pages 36.7.1–36.7.4, 2003.

[9] T.S. Barnett and A.D. Singh. “Relating yield models to burn-in fall-out in time”. *Proc. of Int. Test Conf.*, pages 77–84, 2003.

[10] T.S. Barnett, A.D. Singh, M. Grady, and K.G. Purdy. “Redundancy implications for product reliability: Experimental verification of an integrated yield-reliability model”. *Proc. of Int. Test Conf.*, page 2002, 693-699.

[11] T. Barrette, V. Bhide, K. De, M. Stover, and E. Sugasawara. “Evaluation of Early Failure Screening Methods”. *IEEE IDDQ Workshop*, pages 14–17, 1996.

[12] J. Black. “Electromigration : A brief survey and some recent results”. *IEEE Trans. on Electron Devices*, ED-16(4):338–347, 1969.

[13] D.L. Blackburn and A.R. Hefner. “Thermal components models for electro-thermal network simulation”. *Proc. of 9th IEEE SEMI-THERM Symposium*, pages 88–98, 1993.

[14] M. Borah, R.M. Owens, and M.J. Irwin. “Transistor sizing for low power CMOS circuits”. *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, 15(6):665–671, 1996.

[15] S. Borkar. “Design challenges of technology scaling”. *IEEE Micro*, pages 23–29, July-August 1999.

[16] S. Borkar. "Leakage reduction in digital CMOS circuits". *Proc. of IEEE Solid-State Circuits Conf.*, pages 577–580, 2002.

[17] J.B. Bowles. "A survey of reliability-prediction procedures for microelectronics devices". *IEEE Trans. on Reliability*, 41(1):2–12, 1992.

[18] Y.K. Cheng, C.C. Teng, A. Dharchoudhury, E. Rosenbaum, and S.M. Kang. "A chip-level electrothermal simulator for temperature profile estimation of CMOS VLSI chips". *Proc. of Int. Symposium on Circuit and Systems*, pages 580–583, 1996.

[19] Micro Control Co. http://www.microcontrol.com/.

[20] Siborg Corp. web-site: http://www.siborg.com/.

[21] R. Daasch, K. Cota, J. McNames, and R. Madge. "Neighbor selection for variance reduction in IDDQ and other parametric data". *Int. Test Conf. (ITC)*, pages 92–100, 2001.

[22] B. Davari, R.H. Dennard, and G.G. Shahidi. "CMOS scaling for high performance and low power - The next ten years". *Proc. of the IEEE*, 83(4):595–606, 1995.

[23] N.F. Dean and A. Gupta. "Characterization of a thermal interface material for burn-in application". *Proc. of IEEE Thermal and Thermomechanical Phenomena in Electronic Systems*, pages 36–41, 2000.

[24] J. Van der Pol, F. Kuper, and E. Ooms. "relation between yield and reliability of integrated circuits and application to failure rate assessment and reduction in the one digit fit and ppm reliability era". *Microelectronics and Reliability*, 36(11/12):1603–1610, 1996.

[25] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau. "Transistor element for 30 nm physical gate lengths and beyond". *Intel Tech. Journal*, 6(2):42–54, 2002.

[26] International Technology Roadmap for Semiconductors (ITRS). http://public.itrs.net/.

[27] D.J. Frank. "Power-constrained CMOS scaling limits". *IBM Journal of Research and Development*, 46(2/3):235–244, 2002.

[28] K. Fukasaku, A. Ono, T. Hirai, Y. Yasuda, N. Okada, S. Koyama, T. Tamura, Y. Yamada, T. Nakata, M. Yamana, N. Ikezawa, T. Matsuda, K. Arita, H. Nambu, A. Nishizawa, K. Nakabeppu, and N. Nakamura. "UX6-100 nm generation CMOS integration technology with Cu/Low-k interconnect". *Proc. of Symposium on VLSI Technology*, pages 64 –65, 2002.

[29] D. Gardell. "Temperature control during test and burn-in". *IEEE Inter Society Conf. on Thermal Phenomena*, pages 635–643, 2002.

[30] G. Gerosa, M. Alexander, J. Alvarez, C. Croxton, M. D'Addeo, A.R. Kennedy, C. Nicoletta, J.P. Nissen, R. Philip, P. Reed, H. Sanchez, S.A. Taylor, and B. Burgess. "A 250-MHz 5-W PowerPC microprocessor with on-chip L2 cash controller". *IEEE Journal of Solid-State Circuits*, 32(11):1635–1649, 1997.

[31] T.J. Goh, A.N. Amir, C.P. Chiu, and J. Torresola. "Novel thermal validation metrology based on non-uniform power distribution for Pentium III Xeon cartridge processor design with integrated level two cache". *Proc. of Electronic Components and Technology Conference*, pages 1181 –1186, 2001.

[32] T.J. Goh, K.N. Seetharamu, G.A. Quadir, and Z.A. Zainal. “Thermal methodology for evaluating the performance of microelectronic devices with non-uniform power dissipation”. *Proc. of IEEE Electronics Packaging Technology Conf.*, pages 312–317, 2002.

[33] S.H. Gunter, F. Binns, D.M. Carmean, and J.C. Hall. “Managing the impact of increasing microprocessor power consumption”. *Intel Tech. Journal*, Q1:1–9, 2001. http://developer.intel.com/technology/itj/archive.htm.

[34] H.E. Hamilton. “Thermal aspects of burn-in of high power semiconductor devices”. *IEEE Inter Society Conf. on Thermal Phenomena*, pages 626–634, 2002.

[35] C.F. Hawkins, A. Keshavarzi, and J.M. Soden. “Reliability, test and Iddq measurements”. *IEEE Int. Workshop on Iddq testing*, pages 96–102, 1997.

[36] T. Henry and T. Soo. “Burn-In Elimination of a High Volume Microprocessor Using IDDQ”. *Int. Test Conf. (ITC)*, pages 242–249, 1996.

[37] C.K. Hu, R. Rosengerg, H.S. Rathore, D.B. Nguyen, and B. Agarwala. “Scaling effect on electromigration in on-chip Cu wiring”. *IEEE Int. Interconnect Technology Conf.*, pages 267–269, 1999.

[38] S.F. Huang, C.Y. Lin, Y.S. Huang, T. Schafbauer, M. Eller, Y.C. Cheng, S.M. Cheng, S. Sportouch, W. Jin, N. Rovedo, A. Grassmann, Y. Huang, J. Brighten, C.H. Liu, B.V. Ehrenwall, N. Chen, J. Chen, O.S. Park, and M. Common. “High-performance 50 nm CMOS devices for microprocessors and embedded processor core applications”. *IEDM*, pages 237–240, 2001.

[39] Despatch Industries. http://www.despatch.com/pdfs/PBC.pdf.

[40] R.C. Joy and E.S. Schlig. "Thermal properties of very fast transistors". *IEEE Trans. on Electron Devices*, ED-17(8):586–594, 1970.

[41] A.B. Kahng. "ITRS-2001 design ITWG". *ITRS Release Conf.*

[42] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai. "Design Impact of Positive Temperature dependence on Drain Current in Sub-1-V CMOS VLSIs". *IEEE Journal of Solid-State Circuits*, 36(10), 2001.

[43] R. Kawahara, O. Nakayama, and T. Kurasawa. "The Effectiveness of IDDQ and High Voltage Stress for Burn-in Elimination". *IEEE IDDQ Workshop*, pages 9–14, 1996.

[44] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De. "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs". *Int. Symp. Low Power Electronics and Design (ISLPED)*, pages 207 –212, 2001.

[45] T. Kim and W. Kuo. "Modeling manufacturing yield and reliability". *IEEE Trans. on Semiconductor Manufacturing*, 12(4):485–492, 1999.

[46] T. Kim, W. Kuo, and W.K. Chien. "Burn-in effect on yield". *IEEE Trans. on Electronics Packaging Manufacturing*, 23(4):293–299, 2000.

[47] G. Kromann. "Thermal management of a C4/CBGA interconnect technology for a high-performance RISC microprocessor: The Motorola PowerPC 620TM microprocessor". *Proc. of IEEE Electronic and Tech. Conf.*, pages 652–659, 1996.

[48] P. Lall. "Tutorial: temperature as an input to microelectronics-reliability models". *IEEE Trans. on Reliability*, 45(1):3–9, 1996.

[49] B. Lian, T. Dishongh, D. Pullen, H. Yan, and J. Chen. "Flow network modeling for improving flow distribution of microelectronics burn-in oven". *IEEE Inter Society Conf. on Thermal Phenomena*, pages 78–81, 2000.

[50] W.B. Loh, M.S. Tse, L. Chan, and K.F. Lo. "Wafer-level electromigration reliability test for deep submicron interconnect metallization". *IEEE Hong Kong Electron Device Meeting*, pages 157–160, 1998.

[51] R. Madge. "Screening MinVDD Outliers Using Feed-Forward Voltage Testing". *Int. Test Conf. (ITC)*, pages 673–682, 2002.

[52] R. Mahajan, R. Nair, V. Wakharkan, J. Swan, J. Tang, and G. Vandentop. "Emerging directions for packaging technologies". *Intel Tech. Journal*, 6(2):62–75, 2002. http://developer.intel.com/technology/itj/2002/volume06issue02/.

[53] T. M. Mak. "Is CMOS more reliable with scaling?". *www.intel.com/technology/itj/q11999/articles/art-6who.htm.*

[54] T.M. Mak. "Is CMOS more reliable with scaling?". *IEEE Int. On-Line Testing Workshop*, July 2002.

[55] S. McEuen. "Reliability Benefits of IDDQ". *Journal of Electronic Testing: Theory and Applications (JETTA)*, 3(4):327–335, 1992.

[56] J.W. McPherson, V.K. Reddy, and H.C. Mogul. "Field-enhanced Si-Si bond-breakage mechanism for time-dependent dielectric break-down in thin-film SiO2 dielectrics". *Appl. Phys. Lett.*, 71(8):1101–1103, 1997.

[57] F. Monsieur, E. Vincent, D. Roy, S. Bruyere, G. Pananakakis, and G. Ghibaudo. “Time to breakdown and voltage to breakdown modeling for ultra-thin oxides ($T_{OX} < 32A°$)”. *Proc. of IEEE Int. Reliability Workshop (IRW)*, pages 20–25, 2001.

[58] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan. “Scaling stack effect and its application for leakage reduction”. *Int. Symp. Low Power Electronics and Design (ISLPED)*, pages 195 –200, 2001.

[59] S. Narendra, V. De, S. Borkar, D. A. Antoniadis, and A.P. Chandrakasan. “Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-/spl mu/m CMOS”. *IEEE Journal of Solid-State Circuits*, 39(3):501–510, 2004.

[60] P.E. Nicollian, W.R. Hunter, and J.C. Hu. “Experimental evidence for voltage driven breakdown models in ultra thin gate oxides”. *Proc. of IEEE Int. Reliability Physics Symposium*, pages 7–15, 2000.

[61] P. Nigh, D. Vallett, P. Patel, J. Wright, F. Motika, D. Forlenza, R. Kurtulik, and W. Chong. “Failure analysis of timing and IDDQ-only failures from the SEMATECH test methods experiment”. *Int. Test Conf. (ITC)*, pages 43 –52, 1998.

[62] E. J. Nowak. “Maintaining the benefits of CMOS scaling when scaling bogs down”. *IBM Journal of Research and Development*, 48(2/3), 2002.

[63] E. T. Ogawa, Ki-Don Lee, V. A. Blaschke, and P. S. Ho. “Electromigration Reliability Issues in Dual-Damascene Cu Interconnections”. *IEEE Trans. on reliability*, 51(4):403–419, 2002.

[64] A. Ono, K. Fukasaku, T. Hirai, S. Koyama, M. Makabe, T. Matsuda, M. Takimoto, Y. Kunimune, N. Ikezawa, Y. Yamada, F. Koba, K. Imai, and N. Nakamura. “A 100

nm node CMOS technology for practical SOC application requirement". *IEDM*, pages 511–514, 2001.

[65] A. Poppe, G. Farkas, M. Rencz, Z. Benedek, L. Pohl, V. Szekely, K. Torki, S. Mir, and B. Courtois. "Design issues of a multi-functional intelligent thermal test die". *Proc. of IEEE SEMI-THERM Symp.*, pages 50–56, 2001.

[66] J. M. Rabaey. *"Digital Integrated Circuits"*. Prentice Hall, U.S.A, 1996.

[67] P. Reed, M. Alexander, J. Alvarez, M. Brauer, C.C. Chao, C. Croxton, L. Eisen, T. Le, T. Ngo, C. Nicoletta, H. Sanchez, S. Taylor, N. Vanderschaaf, and G. Gerosa. "A 250-MHz 5-W PowerPC microprocessor with on-chip L2 cash controller". *IEEE Journal of Solid-State Circuits*, 32(11):1635–1649, 1997.

[68] A.W. Righter, C.F. Hawkins, J.M. Soden, and P. Maxwell. "CMOS IC reliability indicators and burn-in economics". *Proc. of Int. Test Conf*, pages 194–203, 1998.

[69] N. Rinaldi. "Thermal analysis of solid-state devices and circuits: an analytical approach". *Solid-State Electronics*, 44(10):1789–1798, 2000.

[70] N. Rinaldi. "On the modeling of the transient thermal behavior of semiconductor devices". *IEEE Trans. on Electron Devices*, 48(12):2796–2802, 2001.

[71] K. Roy. "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits". *Proc. of IEEE*, 91(2):305 –327, 2003.

[72] S. Rusu. "Trends and challenges in VLSI technology scaling toward 100 nm". *ESSCIRC*, 2001. web-page:http://www.esscirc.org/esscirc2001/C01-Presentati.ns/404.pdf.

[73] H. Sanchez, B. Kuttanna, T. Olson, M. Alexander, G. Gerosa, R. Philip, and J. Alvarez. "Thermal management system for high performance PowerPC microprocessors". *Proc. of IEEE COMPCON*, pages 325–330, 1997.

[74] C. Schuermyer, B. Benware, K. Cota, R. Madge, R. Daasch, and L. Ning. "Screening VDSM Outliers Using nominal and Subthreshold Supply Voltage IDDQ". *nternational Test Conf. (ITC)*, pages 565–573, 2003.

[75] J.H. Suehle. "Ultra thin gate oxide reliability: Physical models, statistics, and characterization". *IEEE Trans. on Electron Devices*, 49(6):958–971, 2002.

[76] Aehr Test Systems. www.aehr.com.

[77] S. M. Sze. *"Physics of Semiconductor Device"*. John Wiley & Sons, Inc., U.S.A, 1936.

[78] V. Szekely. "Thermal monitoring of microelectronic structures". *Microelectronic Journal*, 25(3):157–170, 1994.

[79] P. Tadayon. "Thermal challenges during microprocessor testing". *Intel Technology Journal*, Q3:1–8, 2000.

[80] C.C. Teng, Y.K. Cheng, E. Rosenbaum, and S.M. Kang. "ITEM: A temperature-dependent electromigration reliability diagnosis tool". *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 16(8):882–893, 1997.

[81] S. Thompson, P. Packan, and M. Bohr. "MOS scaling: transistor challenges for the 21st century". *Intel Tech. Journal*, Q3:1–19, 1998. http://developer.intel.com/technology/itj/archive.htm.

[82] D.P. Valett and J.M. Soden. "Finding fault with deep-submicron ICs". *IEEE Spectrum*, pages 39–50, October 1997.

[83] A. Vassighi, O. Semenov, and M. Sachdev. “Impact of power dissipation on burn-in test environment for sub-micron technologies”. *Proc. of IEEE Int. Workshop on Yield Optimization and Test*, pages 1–5, 2001.

[84] A. Vassighi, O. Semenov, and M. Sachdev. “Thermal Runaway Avoidance”. *IEEE International Reliability Physics Symposium*, pages 655–656, 2004.

[85] A. Vassighi, O. Semenov, M. Sachdev, and A. Keshavarzi. “Thermal management of high performance microprocessors in burn-in environment”. *Proc of 18th IEEE International Symposium on Defect and Fault tolerance in VLSI Systems*, 2003.

[86] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur. “Thermal performance challenges from silicon to systems”. *Intel Technology Journal*, Q3:1–16, 2000. http://developer.intel.com/technology/itj/archive.htm.

[87] K. Wallquist. “On the Effectiveness of ISSQ Testing in Reducing Early Failure Rate”. *Int. Test Conf. (ITC)*, pages 910–915, 1995.

[88] L. Wei, Z. Chen, K. Roy, M.C. Johnson, Y.Ye, and V.K. De. “Design and optimization of dual-threshold circuits for low-voltage low-power applications”. *IEEE Transactions on VLSI Systems*, 7(1):16–24, 1999.

[89] L. Wei, K. Roy, , and V.K. De. “Low voltage low power CMOS design techniques for deep submicron ICs”. *Int. Conference on VLSI Design*, pages 24–29, 2000.

[90] W. Wondrak. “Physical limits and lifetime limitations of semiconductor devices at high temperature”. *Microelectronics Reliability*, 39(6-7):1113–1120, 1999.

[91] J.W. Worman. “Sub-millisecond thermal impedance and steady state thermal resistance explored”. *Proc. of IEEE SEMI-THERM Symp.*, pages 173–181, 1999.

[92] A.M. Yassine, H.E. Nariman, M. McBride, M. Uzer, and K.R. Olasupo. "Time dependent breakdown of ultra-thin gate oxide". *IEEE Trans. on Electron Devices*, 47(7):1416–1420, 2000.