2

The scaling of MOSFETs, Moore's law, and ITRS

For the last three decades, the microelectronic industry has benefited enormously from the MOSFET miniaturization. The shrinking of transistors to dimensions below 100 nm enables hundreds of millions transistors to be placed on a single chip. The increased functionality and reduced cost of large variety of integrated circuits and systems has brought its own benefit to the end users and above all the semiconductor industry. A low cost of manufacturing, increased speed of data transfer, computer processing power and the ability to accomplish multiple tasks simultaneously are some of the major advantages gained as a result of transistor scaling.

This chapter has four main sections. The first section deals with the Moore's law and its impact on the overall development of semiconductor technology and on MOSFET scaling in particular. Then the contributions made by the International Technology Roadmap for Semiconductors (ITRS) to the advancement of microelectronics technology from the MOSFET point of view are briefly discussed. At the same time the influence of the ITRS on the current priorities and directions related to the scaling of transistors will be discussed.

Section two describes the two basic forms of scaling considered by industry and research communities. Some of the fundamental limitations that that will eventually limit the scaling of conventional MOSFETs are examined in section three. The chapter ends with a summary presented in section four.

2.1 The impact of Moore's law and ITRS on device scaling

Moore's 'law' and the ITRS have been complimenting each other since the first edition of the roadmap in the early 90's. The former has been cast as a law from engineering observation made by G. Moore in the mid sixties [2:1]. It was initially a forecast on the number of transistors that can be integrated into a microchip for the next ten years (1965-1975), but the trend remained almost unchanged over the next three decades. The ITRS on the other hand is a comprehensive guide that enables the semiconductor industry to transform this observation into reality. At this stage, however, one has to be careful when interpreting "Moore's law", as a physical or mathematical law. Despite the efforts made by Meindle [2.2] to formulate the "compact mathematical formulation of the Moore's law" ($N = F^{-2}D^2P_E$ where N is the number of transistors per chip, F is the minimum feature size, D is the chip area, and P_E is transistor packaging efficiency measured per minimum feature area), it remains simply an empirical observation on the rate of growth of semiconductor technology [2.3] originating from the forecast depicted inset to figure 2:1. Therefore, in order to clarify its role on the growth of semiconductor industry, in the following sub sections Moore's law is discussed briefly together with the ITRS mainly from the MOSFET scaling point of view.

2.1.1 Moore's Law

Back in time when Gordon Moore published his article, "Cramming more components onto integrated circuits" in 1965 [2:1], he was probably not aware of its impact on the remarkable progress of semiconductor technology in the years to come. In this publication he made an observation that it will be possible to integrate 6.5×10^4 components into a single chip by 1975, provided that the number of active transistor per chip doubled roughly every year. As illustrated in figure 2:1 the advances of the semiconductor technology have been able to follow this predicted trend.

When G. Moore made his prediction, the number of transistors in a single chip was roughly 32 and today there are approximately half a billion transistors integrated on a single microprocessor (figure 2:1). This phenomenal growth has demonstrated how

visionary his prediction was, and how vital has it been to the technology enabling the shrinking of individual transistors. The scaling of MOSFETs, which are the key components in digital technology, has revolutionized the semiconductor industry and has also enabled the realization of the immensely complex devices and systems we rely on at present.

Although the "Moore's law" has been interpreted differently at the different stages of the semiconductor technology industry's development, the formulation that has been accepted as a general consensus states that: "the number of components per chip doubles every 18 months" [2.4]. Note that the original assumption made by Moore, according to the inset in figure 2.1, was that the number of components per chip will be doubled every 12 months. Indeed the originally stated rate of development was maintained in the seventies, as shown by Moore himself in 1975 [2.5], and continued to the early eighties. The present 18 months period of doubling of the chip components is a modification in line with the past and present (2003) ITRS editions and the real state of the industry.

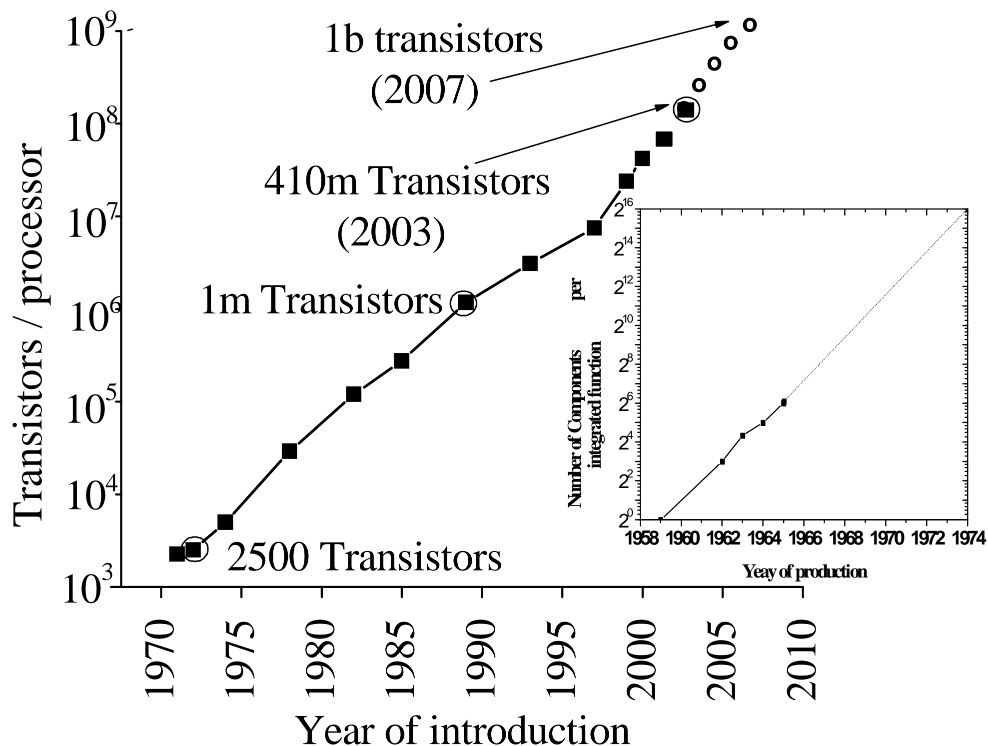

Figure 2:1 Visualization of Moore's Law: The number of transistors integrated in a commercially available processor and the outlook towards a billion transistors in a single processor due in year 2007 (Intel). The inset graph: Projection made by Moore on his original paper on the number of components per integrated device [2.1]

2.1.2 Implication of Moore’s law

Moore’s law has had various implications on the microelectronics manufacturing industry and user applications in general over the last 30 years. As a result, increasing functionality [2.6], cost per function reduction, and better performance, have all been achieved for every new generation of integrated circuits.

According to the ITRS, the functionality is defined as the number of bits in a DRAM chip or the number of logic transistors in a microprocessor unit. With the integration of more individual components in a single chip the functionality per chip increases (figure 2:2) together with the increase in the density of functions (functions/area). The increase in functionality minimizes the delay of data flow that occurs due to the isolation of individual functions on separately integrated systems [2.7]. More functionality also means an increase in overall physical density of useable transistors per total chip area. Figure 2:2 shows according to the ITRS, that in both the near-term (2003-2009) and the long-term (20010-2018) functionality will be increasing by roughly 100% in every technology nodes.

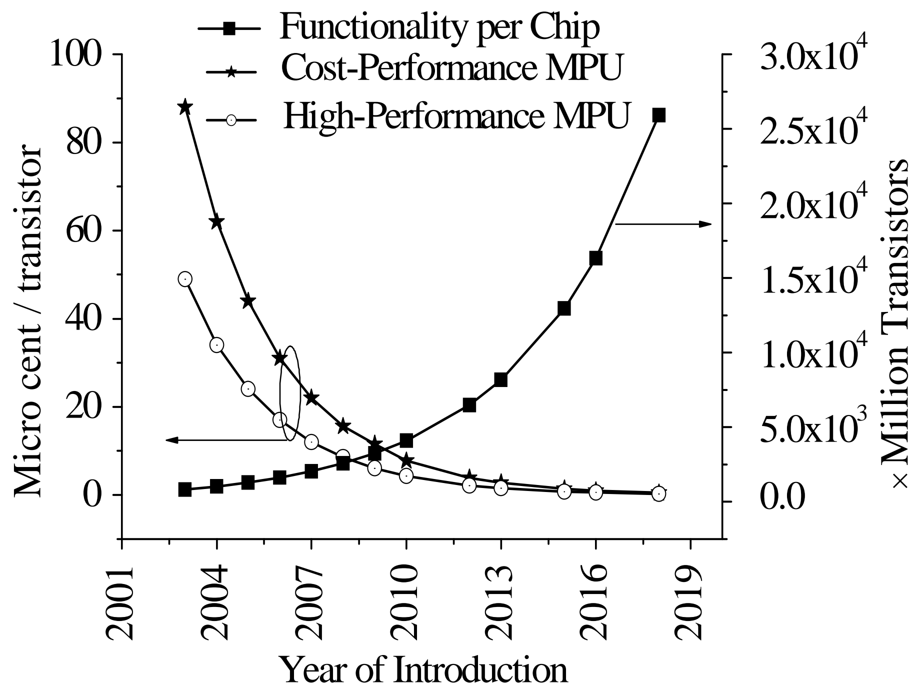

Figure 2:2 Cost – performance of Microprocessor Unit (MPU), cost of high speed performance MPU and functionality (functionality is often associated with the number of bits or unit devices in MPU) against the year of introduction of technology nodes: data ITRS 2004 update.

The second important feature associated with the Moore's law is cost. It is a general rule that the goal of every manufacturing community is maximizing the profit while minimizing the cost of production. The electronics industry is not unique in this. In fact the primary implication of the Moore's law is the reduction of manufacturing cost per function and at the same time to increase the functionality per chip. As it can be seen from figure 2:2 the reduction in cost-per-function according to the latest ITRS edition, is roughly 50% in about two years.

The third important implication of Moore's law is the performance factor. Performance in general can be measured, for example, by the speed of typical microprocessors. Figure 2:3 shows the increasing speed and density of present and future generations of technology nodes. The off-chip frequency is the maximum input and output signal frequency to board peripheral buses of high performance devices [2.4]. The off-chip frequency is increasing faster than the on-chip local frequency near the end of the current edition of the ITRS.

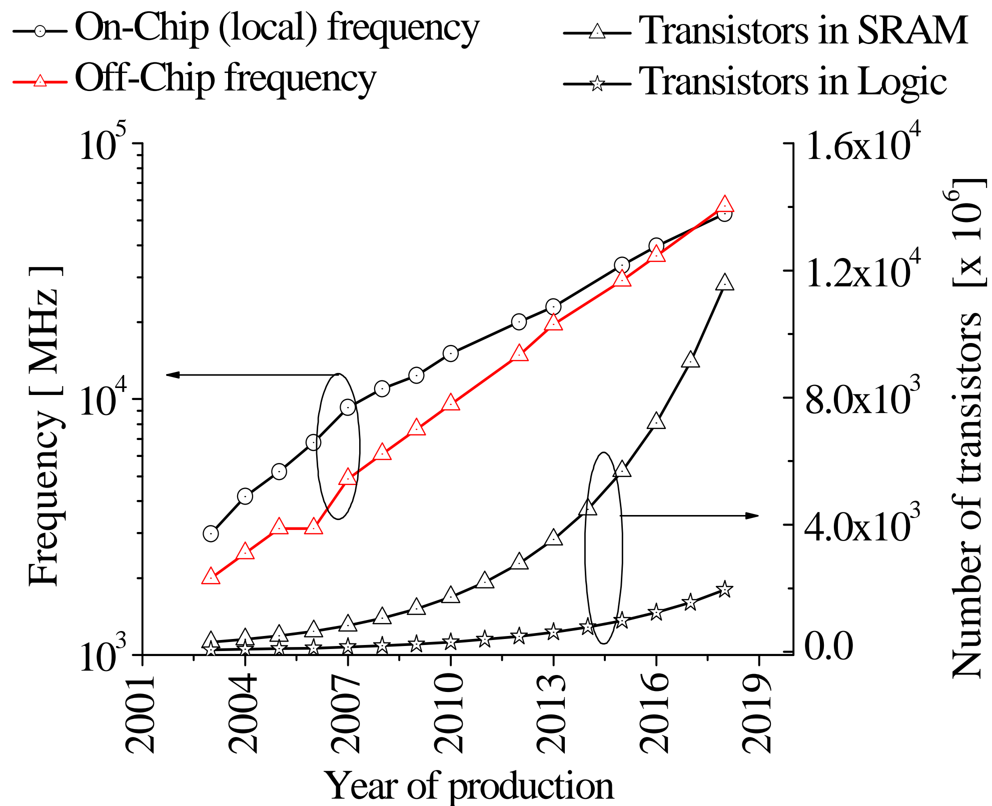

Figure 2:3 The technology trends of on-chip local clock, off-chip frequency, chip density in SRAM and Logic gats (transistors per cm^2). ITRS 2004 Edition

In general the frequency is related to the speed of switching of the individual and simple logic components determined by CMOS transistors which is inversely proportional to the delay time that takes to propagate signal through the inverter. The inverter delay time formulated as [2.8]:

$$\tau_{\text{int}} = R_{sw}(C_{in} + C_{out}) \tag{2.1}$$

Where τ_{int} is the inverter delay time, R_{sw} is the switching resistance, C_{in} input capacitance, C_{out} the output capacitance, and in equation 2.2, C_{gate} is the gate capacitance. This inverter delay time can be used as approximation of the CMOS delay time which is calculated empirically as: [2.9]

$$\tau = C_{gate}\frac{V_{DD}}{I_{dsat}} \tag{2.2}$$

According to the generalised scaling [2.10, 2.11], τ is inversely proportional to the scaling factor, which allows faster circuit operations. Figure 2:2 also depicts the increase in the density of transistors in SRAM and Logic circuits. Density is inversely proportional to the total chip area (*1/A*). Therefore, density will increase by κ^2 as a result of scaling, where $\kappa \approx \sqrt{2}$ is the scaling constant (see section 2.1.4).

2.1.3 The International Technology Roadmap for Semiconductors

The technology roadmap is an ambitious document widely used as a guiding reference for advanced semiconductor device research and manufacturing purposes. The latest edition (2003)[†] of international technology roadmap for semiconductors (ITRS), updated in 2004, sets main objectives and targets to 2018. Based on research from the semiconductor industry and academia, the latest edition of the ITRS outlines the requirements and identifies the challenges which allow Moore's law to be maintained over the next 15 years. In addition to the challenges, it also outlines the possible solutions to some of the problems that the industry may face and highlights the specific areas that need urgent research.

Overall, the roadmap has three major contributions. The first is to identify the needs and requirements to be met by technology solutions currently under development. The second is to recognize the existence of interim solutions for the medium term challenges and problems and their limitations at the present time. The third important

[†] The latest ITRS edition of 2005 has been published after the completion of this thesis

contribution of ITRS is to identify the areas where there are "no known manufacturing solutions" customarily labelled as the "Red Brick Wall" - to induce the industry to concentrate on them strategically and focus research efforts in these areas.

The ITRS is a comprehensive document with more than 600 pages, including the executive summary. It covers fifteen categories related to semiconductor industry and to basic research and development areas. One of the important sections expanded significantly in the latest edition ITRS is on the emerging research devices. It was organized with the aim of finding and building successful new device structures that can replace conventional MOSFETs. Although some of the listed structures are more of research type, the device structures such as fully depleted silicon on insulator (FD SOI) and the multiple gates architectures including the double gate MOSFETs and FinFETs are the promising candidates to replace mainstream device structures.

Year of Production	2003	2004	2005	2006	2007	2008	2009
Technology Node		hp90			hp65		
DRAM ½ *Pitch (nm)*	100	90	80	70	65	57	50
MPU/ASIC M1 ½ *Pitch (nm)*	120	107	95	85	75	67	60
MPU/ASIC *Poli Si ½ Pitch (nm)*	100	90	80	70	65	57	50
MPU Printed L_g (nm)	65	53	45	40	35	32	28
MPU Physical L_g (nm)	45	37	32	28	25	22	20
Equivalent t_{ox} (nm)	1.3	1.2	1.1	1	0.9	0.8	0.8
V_{dd} (HP)	1.2	1.2	1.1	1.1	1.1	1	1
Off current , I_{off} [μA/μm]	0.03	0.05	0.05	0.05	0.07	0.07	0.07
Drive current, I_{on} [μA/μm]	980	1110	1090	1170	1510	1530	1590
HP NMOS intrinsic delay τ [ps]	1.2	0.95	0.86	0.75	0.64	0.54	0.48
Relative intrinsic speed, 1/τ	1	1.26	1.39	1.6	1.86	2.2	2.49
Logic gate delay [ps]	30.24	23.94	21.72	18.92	16.23	13.72	12.13
DRAM cell size [μm^2]	0.082	0.065	0.048	0.036	0.028	0.019	0.015
S/D extension x_j [nm]	24.8	20.4	17.6	15.4	13.8	8.8	8.0

Table 2:1 The near term years (2003-2009) of selected overall roadmap technology characteristics that are required to continue the present scaling trends of conventional MOSFETs. Half pitch 90 and 65nm technology nodes are marked as hp90 and hp65 respectively. (ITRS 2003 edition)

However, not all of these categories and data are relevant to this research. Therefore, in this sub-section we only concentrate on the high performance devices, which are in the heart of this work. The summarised data of device dimensions and electrical parameters for high performance devices depicted in tables 2:1 (near-term years) and 2:2 (long-term years) have been adopted as a guide for the scaling of the 35 nm MOSFET described in chapter 4.

The carefully calibrated 35 nm gate length MOSFETs manufactured by Toshiba [2.12]) were used as a basis for further scaling to gate lengths of 25, 18, 13, and 9 nm transistors. The overall calibration and scaling methodology and results are presented in chapter 3 and 4 respectively. The dimensions of the 35 nm MOSFET physical gate length used for this work are not characteristics of particular node on the ITRS roadmap. It's performance, I_{on} = 676μA/μm, I_{off} = 100nA at V_{dd} = 850mV and design parameters, t_{ox}= 1.2 nm x_j = 20 nm are close to the 37 nm high performance device required for the 90 nm node and 80 nm technology generations.

Year of Production Technology Node	2010 hp45	2012	2013 hp32	2015	2016 hp22	2018
DRAM ½ Pitch (nm)	45	35	32	25	22	18
MPU/ASIC M1 *½ Pitch (nm)*	54	42	38	30	27	21
MPU/ASIC *Poli Si ½ Pitch (nm)*	45	35	32	25	22	18
MPU Printed L_g (nm)	25	20	18	14	13	10
MPU Physical L_g (nm)	18	14	13	10	9	7
Equivalent t_{ox} (nm)	0.7	0.7	0.6	0.6	0.5	0.5
V_{dd} (HP) (V)	1	0.9	0.9	0.8	0.8	0.5
Off current , I_{off} [μA/μm]	0.1	0.1	0.3	0.3	0.5	0.5
Drive Current, I_{on} [μA/μm]	1900	1790	2050	2110	2400	2190
HP NMOS intrinsic delay τ [ps]	0.39	0.3	0.26	0.18	0.15	0.11
Relative intrinsic speed, $1/\tau$	3.06	4.05	4.64	6.8	8.08	10.77
Logic gate delay [ps]	9.88	7.47	6.55	4.45	3.74	2.81
DRAM cell size [μm_2]	0.1222	0.0077	0.0061	0.0038	0.0025	0.0016
S/D extension depth x_j [nm] [‡]	7.2	11.2	10.4	8.0	7.2	5.1

Table 2:2 The long - term years (2010-2018)

[‡] The extension depth (x_j) is calculated with the assumption of introducing new device structures beyond year 2007, like fully depleted SOI and multi gate device structures. (ITRS 2003 edition)

Although the electronics industry prefers to continue as long as possible with the scaling of conventional MOSFETs, there is “Red Brick Wall” to this process unless there is a major technological breakthrough. High channel doping, which degrades the device performance, and ultra thin gate oxides, which introduce unacceptable gate leakage, are likely to prompt a replacement to conventional MOSFETs somewhere beyond the 65nm technology node. Among the replacement candidates are, for example, ultra-thin body SOI and multiple gate devices complimented by the introduction of strained silicon in the channel region to enhance the carrier mobility, and high permittivity materials in the gate stack in order to suppress gate leakage. Some of the critical scaling limitation factors will be examined more closely in the next sections of this chapter.

2.1.4 The scaling factors and technology trends

The scaling factor of $\kappa \approx \sqrt{2}$, related to a 70% size reduction of the major technology nodes every two years, has been adapted for the linear scaling of device dimensions in this work. The other scaling constant, α for the electric field and potential used in the generalised scaling scenario [2:10] is not specified on the road map. However, it can be calculated from the supply voltages (V_{dd}), which are specified in the roadmap for corresponding feature sizes and the linear scaling factor κ as:

$$V' = \frac{\alpha}{\kappa} V \Rightarrow \alpha = \kappa \frac{V'}{V} \quad (2.3)$$

V' is the new supply voltage given in the technology roadmap and V is the supply voltage of the previous generation. It should be noted that in some papers [2.13], the linear scaling factor has been decomposed to separate dimensional scaling parameters in the so called “selective scaling case”, which introduces different values for vertical, horizontal, and lateral dimensions multipliers. However, in this work, the generalized scaling rule has been adopted as a principal guiding rule for device scaling. A review of the different scaling approaches is presented in the next section.

Unlike the previous editions of ITRS, no prediction of the technology acceleration has been made in its latest edition (ITRS'03). Also, as illustrated in figure 2.4, in the last ITRS edition, the technology generations are predicted to shift from the present two-year cycle to a three-year cycle trend around 2007. The technology node

continued to be defined as 70% dimension reduction per node or approximately 50% reduction per two nodes. The "technology-node-cycle" is the period of time in which a new technology node is introduced.

In addition to the scaling of the gate length, the oxide thickness is another critical parameter, which has been aggressively scaled down in order to achieve a sufficient drive current and to control short channel effect. The later can be achieved by maintaining the electrostatic control of the channel potential by the gate.

Figure 2:4 shows the technology half pitch (hp) and gate length trends adopted in the ITRS'03 edition. Beyond year 2007 the two year cycle delays by another year and is expected to be three years until the end the present roadmap projection time-line and probably beyond. The physical gate length is conventionally adapted as minimum feature size regarding the individual devices.

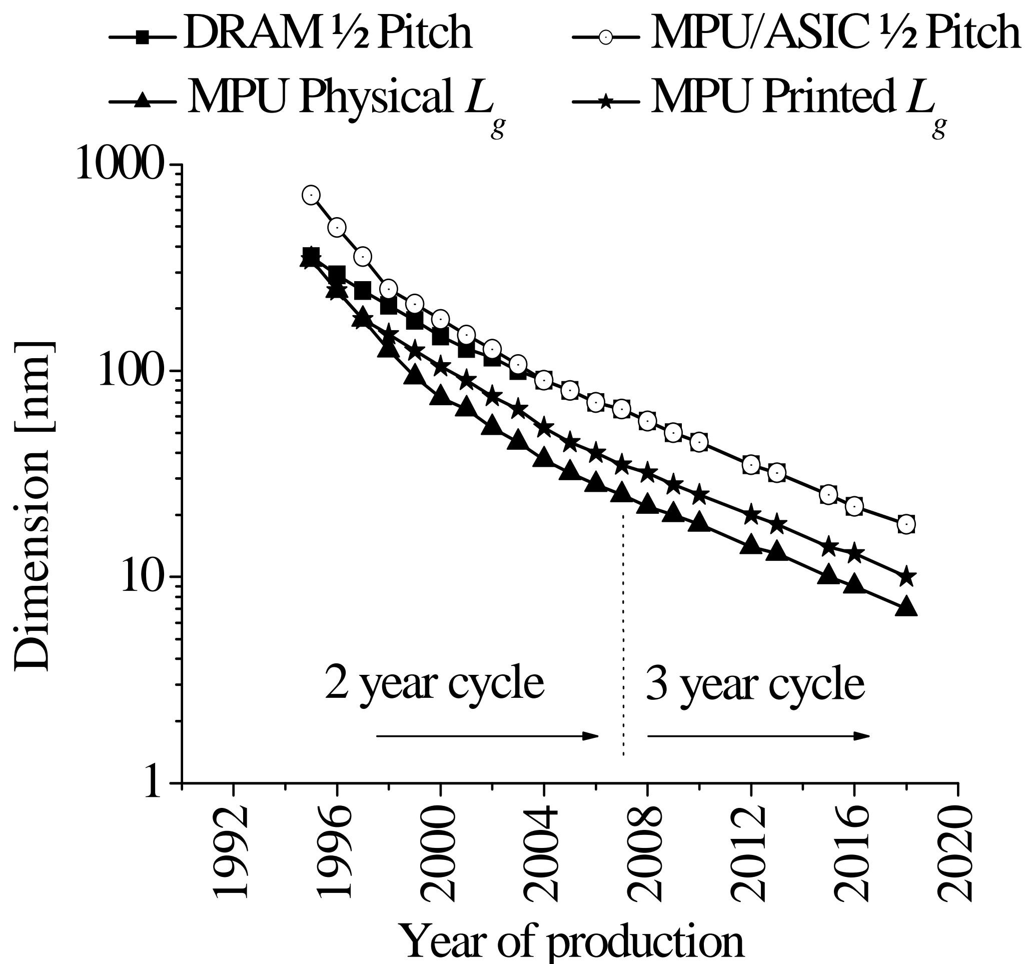

Figure 2:4 Technology half-pitch and gate length trends**.**

2.2 The scaling rules for conventional MOSFETs

In the preceding sections the technology roadmap and the Moore's law have been discussed in order to examine their role in pursuing transistor scaling, and above all in highlighting scaling's unprecedented contributions to the enormous advance of semiconductor technology. Without the extraordinary miniaturization of transistors, it would be impossible to produce higher volumes of faster devices operating at lower power. Nobody in the semiconductor industry disputes this state of affairs. This section further introduces the theory and practice of the scaling process. It begins by reviewing some of the classic papers on the constant field and generalised device scaling rules, followed by detailed analysis of advantages and shortcomings of both rules.

2.2.1 Constant field scaling

Dennard *at al.* presented their pioneering research work on the scaling of MOSFET devices at the International Electron Device Meeting (IEDM) 1972 [2.14] and published a comprehensive paper on the scaling of MOS transistors in 1974 [2.15], from which the "constant field scaling" theory has emerged. The basic principle which they employ is that in order to increase the performance of a MOSFET we must reduce linearly the size of the transistor, together with the supply voltage, and increase the doping concentration in a way which keeps the electric field in the device constant - hence the name "constant field scaling" (figure 2.5).

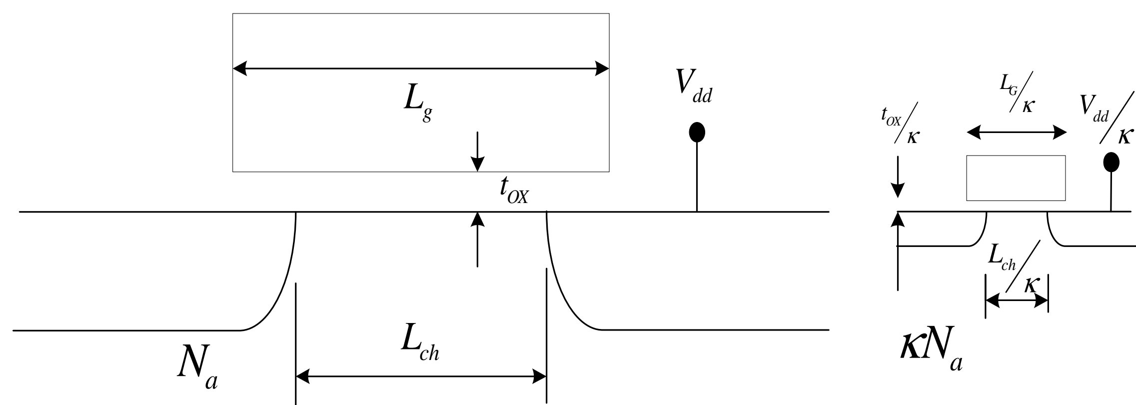

Figure 2: 5 Illustration of MOSFET miniaturisation. The sketch on the right hand is the scaled device according to the constant field rule. (Reference [2.15])

The scaling process is performed by a linear transformation of three design parameters (voltage, doping concentration, and physical dimensions) of a particular generation of transistor by the same scaling factor, κ. The scaled down device will have a reduced voltage (V_{dd}/κ), vertical (t_{ox}/κ and X_j/κ) and horizontal (L/κ) dimensions, and an increased doping concentration (κN_a) as depicted in figure 2.5. Despite the change in those parameters, the intensity of the electric field ($v/L = v'/L'$) remains virtually unchanged since both the dimension and the supply voltage scale by the same ratio.

Table 2.3 summarises the changes in device dimensions and circuit parameters as a result of both the constant field and the generalised scaling rules. Only the most important design and operational parameters are included in the table. The other quantities which are not given in table 2.3 can be deduced using the listed design parameters.

Scaled Parameters	Constant field Scaling	Generalised scaling
t_{OX}, L, W, X_j, W_d	$1/\kappa$	$1/\kappa$
N_a, N_d $(ions/cm^3)$	κ	$\alpha\kappa$
Power supply: (V_{dd})	$1/\kappa$	α/κ
Electric field in device: (E)	1	α
Capacitance: (C)	$1/\kappa$	$1/\kappa$
Inversion charge density (Q)	1	α
Circuit delay time: τ CV/I	$1/\kappa$	$1/\kappa$
Power dissipation: (P)	$1/\kappa^2$	α^2/κ^2
Power density $(\sim P/A)$	1	α^2
Circuit density	κ^2	κ^2
Chip Area (A)	$1/\kappa^2$	$1/\kappa^2$
Current, Drift: (I)	$1/\kappa$	$1/\kappa$

Table 2:3 Summary of the constant field scaling and the generalised scaling rules

The other issue addressed in [2.15] was the application of ion implantation in the fabrication process of the scaled device. This is an important process step which allows us to more accurately place the dopants in the shallower source drain junctions and channel of the scaled device. In addition to its advantages, the shortcomings of the scaling process have also been addressed in [2.15] For example, carrier mobility degradation as a result of high doping in the channel and the short channel effect (decreasing of V_T) are some of the main drawbacks. The adverse effect of doping concentration on carrier mobility can be observed from the empirical formula given by equations (2.4 & 2.5) [2.16].

$$\mu_e = 88T_n^{-0.57} + \frac{7.4\times10^8 T_r^{-2.23}}{1+\left[N_d \big/ 1.26\times10^{17} T_n^{2.4}\right] 0.88T_n^{-0.146}} \tag{2.4}$$

$$\mu_p = 54.3T_p^{-0.57} + \frac{1.36\times10^8 T_r^{-2.23}}{1+\left[N_a \big/ 2.35\times10^{17} T_p^{2.4}\right] 0.88T_p^{-0.146}} \tag{2.5}$$

Where, T_n and T_p are the electron and hole temperatures of interest respectively and T_r is the room temperature. μ_n and μ_p are the electron and hole mobility respectively. N_a and N_d are acceptor and donor concentration. Since carrier mobility is inversely proportional to the channel doping as shown in the equations, increasing channel doping reduces the mobility and the device performance. Regardless of implementation of different channel engineering techniques, the highly doped channel imposes carrier transport problems in aggressively scaled MOSFETs.

2.2.2 Generalised scaling rule

As device dimensions enter into the sub-micron dimensions, two-dimensional effects (short channel effect-SC and drain induced barrier lowering-DIBL) become increasingly important. The gradual field approximation becomes invalid and the field changes significantly even if the constant field scaling scenario is applied. The field also increases due to a much slower reduction in the supply voltage in real circuits compared to the requirements of constant field scaling. This challenge to constant field scaling has

been addressed by Brews *et al.* [2:11], and Baccarani *et al*. [2:10], who have introduced a more generalised scaling theory.

Brews *et al*. mainly concentrated on the minimum channel length for which the subthreshold characteristics of the long channel device can be maintained in the scaled devices. For this purpose they suggested an empirical formula given by:

$$L_{\min} = A\left[x_j t_{ox}(W_s + W_d)^2\right]^{\frac{1}{3}} \qquad (2:6)$$

Where $L_{\min}$ is the minimum channel length, W_s and W_d are the depletion widths in the source and drain regions respectively, x_j is the junction depth, A and t_{ox} are proportionality constant and oxide thickness respectively.

The main advantage of this approach over the constant field scaling, according to [2.11], is that the parameters do not all have to be scaled by one factor. But there is a drawback associated with the way in which the minimum channel length is determined. It was suggested that the channel length could be reduced until a 10% increase in drain current is obtained.

However, the tolerance to the short channel effects may not only depend on a predetermined drain current value, but also depend on circuit applications [2.17]. In addition to this, the scaling of a MOSFET includes to at least five major design parameters (L_g, t_{ox}, V_{dd}, N_a, x_j) [2.8] [2.18], not just the three defined in equation (2.3). The typical electrical behaviour of device under the influence of short channel effect (SCE) and drain induced barrier lowering (DIBL) are highly dependant on all five parameters as shown in equation (2.7 and 2.8)[§] [2.19] [2.16].

$$SCE = 0.64\frac{\varepsilon_{si}}{\varepsilon_{ox}}\left(1+\frac{x_j^2}{L_{el}^2}\right)\frac{t_{ox}}{L_{el}}\frac{W_{dm}}{L_{el}}V_{bi} \qquad (2.7)$$

$$DIBL = 0.80\frac{\varepsilon_{si}}{\varepsilon_{ox}}\left(1+\frac{x_j^2}{L_{el}^2}\right)\frac{t_{ox}}{L_{el}}\frac{W_{dm}}{L_{el}}V_{ds} \qquad (2.8)$$

[§] For further reference on the derivations of equations 2.7 and 2 .8, please look in [2.18]

V_{bi} and V_{ds} are the built in potential and the in put drain Voltage. The effective channel length is defined as $L_{el} = L_g - \Delta L$, where L_g is the physical gate length and ΔL is the sub-diffusion length. It is also clear that from the empirical formulas of CMOS design rules, which require $x_j/L \approx 0.33$, $t_{ox}/L \approx 1/30$ $W_d/L \approx 0.33$, $V_T/V_{dd} \approx 0.20$, that all the five parameters influence the electrostatic integrity of the scaled devices, which determines both SCE and DIBL.

On the other hand, by identifying a significant difference in the two dimensional pattern of the electric field in the active region of the original and the scaled device, Baccarani *et al*. have suggested that the supply voltage and the doping concentration should be scaled with different scaling factor. To facilitate this a new scaling concept an additional scaling constant, α, has been suggested. The effect of α can be demonstrated by examining the Poisson equation within the depletion region, which is explained in [2:8] and given by the equation:

$$\frac{\partial^2(\alpha\psi/\kappa)}{\partial(x/\kappa)^2} + \frac{\partial^2(\alpha\psi/\kappa)}{\partial(y/\kappa)^2} = \frac{qN_a^{'}}{\varepsilon_{si}} \tag{2.8}$$

Where $N_a^{'} = \alpha\kappa N_a$, is the channel doping concentration in the scaled device. Equation (2.8) is based on the assumption that the potential will be scaled by α/κ and the electric field by just α (where $\alpha \geq 1$). It is, however, important to note that in [2.17] the scaled potential is given as $\psi' = \psi/\kappa$ which is different than the one shown in equation (2.3). Moreover scaling the potential by the same factor as the dimensions leads to the constant field scaling theory. The main reason for adopting in this work the generalised scaling with α determined by equation (2.3) is the fact that the supply voltage can not be scaled as fast as the device dimensions due to the non-scaling property of the threshold voltage and the subthreshold slope [2.10].

The main problem with the generalized scaling rule, particularly in deep sub-100 nm scaled devices is an increase on power density ($P/A \Rightarrow \alpha^2$). The scaling of total area scales as $1/\kappa^2$ while the power dissipation per circuit scales as α^2/κ^2, i.e., the size of the area scales down faster than the power dissipation. This difficulty is considered to be one of the major scaling limitation factors [2.20]. The scaling limits are discussed in detail in later sections of this chapter.

2.2.3 Evolution of CMOS design

Although the history of semiconductor devices goes back to the 1920s, this thesis only concentrates on the period after 1960s where the transformation of metal oxide semiconductor field effect transistors (MOSFET) from research devices to the major building block of commercial integrated circuits takes place. For further reading, a comprehensive historical review is presented in [2.21].

The first working CMOS circuit was developed circa 1964 at RCA [2.22] and was integrated in to logic gates in late 1960's [2.23]. The fabrication of both n-MOS and *p*-MOS transistors on the same wafer was an important stage in revolutionizing the integrated circuit. It has the advantages over using single *n* and *p*-MOSFET devices discussed next.

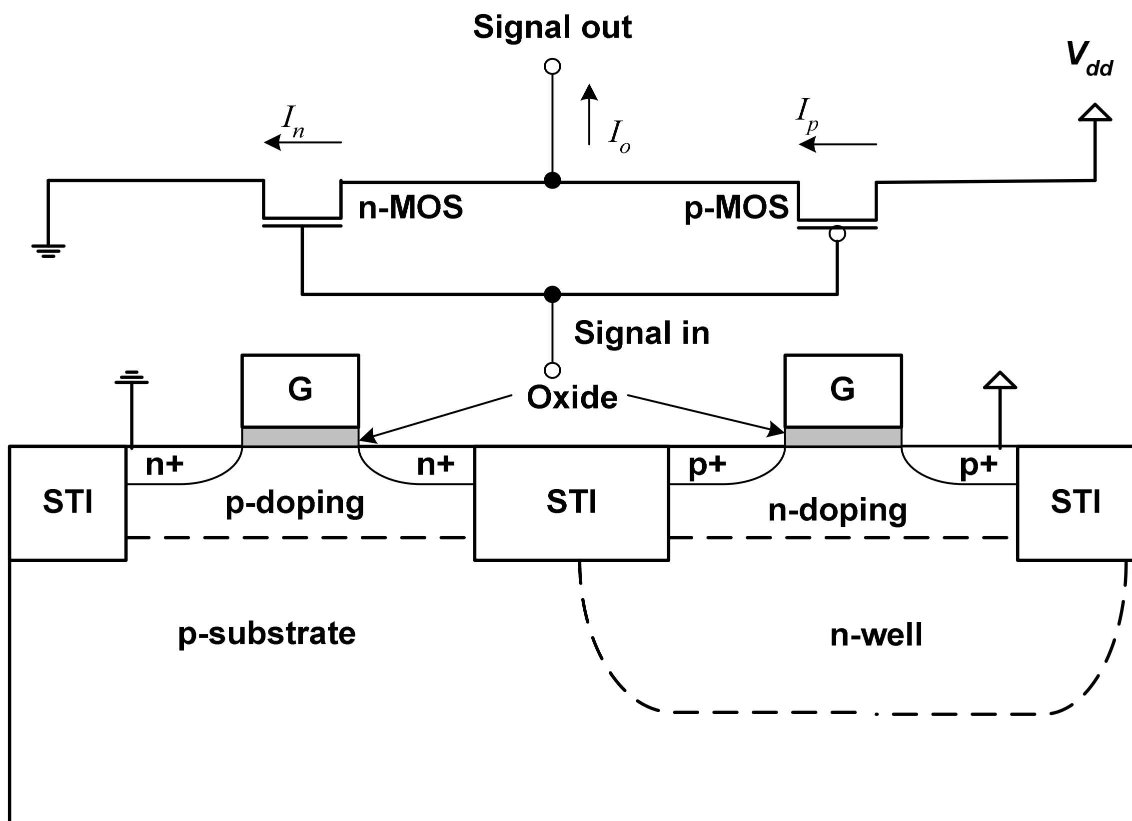

Figure 2:6 Cross-sections of fabricated CMOS device which can be used as inverter circuit. STI stands for shallow trench isolation and the currents I_n and I_p are drain currents of n and p-type devices and I_o is an output current of an inverter.

CMOS circuits offer high switching speed, high density of integration, and very low static power dissipation. These advantages favour the miniaturisation of MOSFETs and subsequent realization of high density integrated circuits with ever increasing speed. One of the simplest building blocks for CMOS logic gates is the inverter illustrated in figure 2:6 which is realized on a single wafer using single, double or triple well technology.

The next important milestone in the evolution of MOSFET design is the introduction of the self-aligned polycrystalline-silicon (poly) gate in the early 80's. The self-alignment of the source and drain to the gate reduces stray capacitance which improves the signal propagation delay, $\tau \approx CV/I$ and overall circuit performance. Moreover polycrystalline-silicon as a material is stable and completely compatible with silicon technology [2.24]. The gate material work function must be suitable in order for the device to have an acceptable subthreshold voltage. Polysilicon has properties which match all these requirements.

As MOSFET channel lengths approach sub-micron dimensions the high electric field in the channel start to affect the device reliability and the introduction of lightly doped drain (LDD) MOSFETs in of late 70's are required [2.25] (see figure 2:7A). The lightly doped n^- region in the neighbourhood of the conventional n^+ source and drain areas soften the electric field and reduce hot carrier injection in the oxide [2.26] [2.27]. With the reduction of drain voltages the need for LDD subsides.

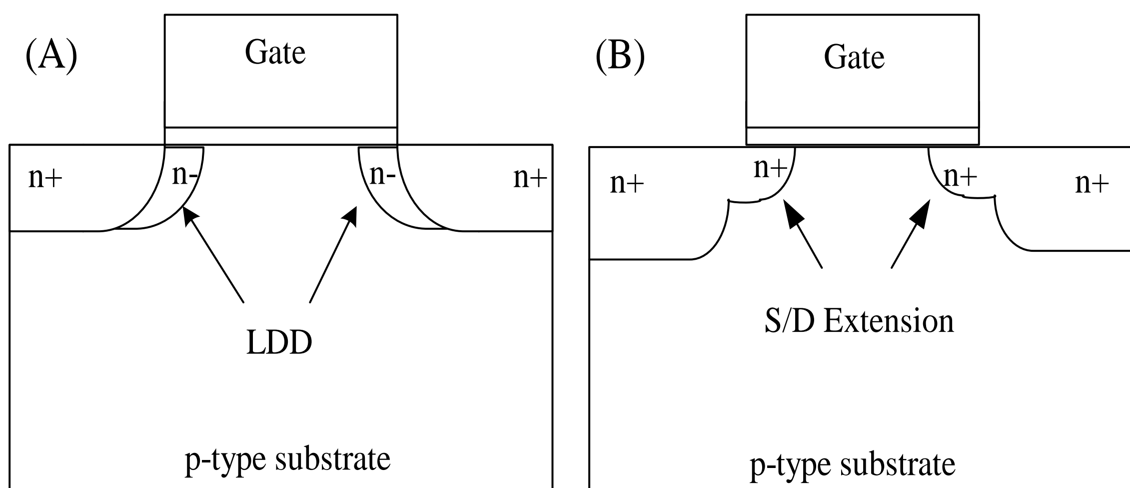

Figure 2:7 Cross section view of LDD n-MOSFET (A) and MOSFET with heavily doped source and drain extensions (B)

Later on, heavily doped but very shallow source and drain extensions (illustrated in figure 2:7B), were introduced to combat short channel effects without introducing problems associated with high series resistance [2.28].

One of the important technology stages during the process optimization and performance enhancement of the individual MOSFET is the introduction of Salicide (Self aligned) into the CMOS production process (see figure 2:8). The increase of the resistance of the poly-gate and the parasitic resistance of the shallow source and drain junction results in a poor performance of the scaled down MOSFETs. Self-aligned silicide technology has been first suggested by Crowder *at al* to be used as a shunting gate electrode on top of poly-gate [2.29].

In late 80's its application expanded to the source and drain electrodes to reduce the access and the contact resistance. Since the sheet resistance increases with the reduction of junction depth, it becomes important to use silicide to reduce the parasitic source and drain resistance in order to achieve the required drive current. $TiSi_2$ and $CoSi_2$ are the most widely used metal silicide in semiconductor industry to day, with the resistivity of 13-16 and 22-28 Ω-cm respectively. $NiSi_2$ is the next promising candidate.

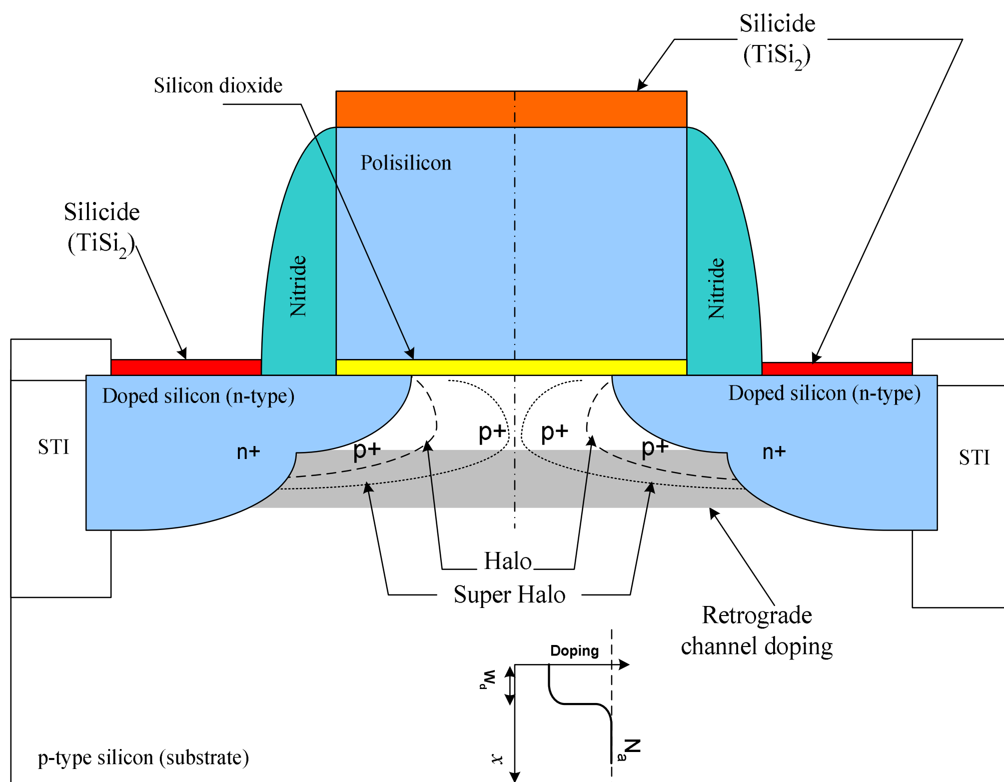

Figure 2:8 schematic illustrations showing self-aligned silicide in source/drain and on the poly-gate, halo (super-halo) doping and retrograde channel doping in a cross sectional view of the MOSFET. The inset depicts the ideal low-high (retrograde) doping profile.

Another key aspect in the evolution of CMOS design is channel engineering. Although various doping schemes have been implemented throughout the history of CMOS technology development, only the retrograde and the halo channel doping which are adopted in the deep submicron MOSFETs designs will be reviewed here.

According to the generalised scaling rule, the channel doping increases $\kappa\alpha$ times in every successive technology generation, in order to control short channel effects. However, this increase in channel doping severely degrades carrier mobility. The traditional channel doping scheme which uses the flat well profile to control V_T and combat short channel effects is not applicable in 100 nm channel length MOSFETs. To overcome problems associated with device performance and to provide relatively lower threshold voltage the super steep retrograde (SSR) channel doping scheme was proposed in early 90's in [2.30]. SSR doping profile increases current drive by enhancing carrier mobility and reducing V_T. [2.31].

For sub 0.1micron transistors the SSR alone is not sufficient to improve the performance (I_{on}) and to control the short channel effect at the same time. To reduce this problem the halo (pocket) doping and latter the supper halo doping [2.32] were introduced (figure 2:8). Halo implantation is performed through larger tilt angles often between 25° and 45°. The pockets are in close proximity to the source and drain controlling the surrounding depletion layers and the SCE, DIBL and the punch through without increasing the channel doping and degrading the device performance.

2.2.4 The present state and the future trends

Conventional CMOS transistors with a gate length of 50 nm and with strained silicon channels [5:33] have already been manufactured successfully, and have been integrated into commercial products [5:34]. According to the technology roadmap, these transistors meet the requirements of the 90 nm technology generation. At the same time the 35 nm MOSFET published by Toshiba and discussed in more detail in chapter four, can also be used for the late stages of 90 nm technology node and the transitional 80 nm inter-node technology.

25 nm gate length conventional MOSFETs, required for the 65 nm technology node, have also been reported in [2:35] It has good subthreshold characteristics (S = 118 mV/decade, I_{off} = 100 nA/µm) and a high drive current (I_{on} = 840µA/µm). Such 25 nm transistors are expected to deliver the performance required by the ITRS. However there

is still a need for '*total process optimization*' [2:35] in order to use the current fabrication technology. For example, better designed doping profiles in the channel, improved gate patterning (minimum LER), shallow junction formation to reduce source-drain sheet resistance, and optimization of the gate oxide (oxynitride) by controlling the amount of nitrogen, are some of the process-optimization steps that can be exploited. In addition to process optimization, the introduction of strained silicon in the channel to enhance carrier transport, and introduction of high-κ materials to replace silicon dioxide as an insulator will improve the chances of using conventional MOSFETs at the 65 nm technology node and possibly beyond.

Apart from the 25 nm gate length MOSFET, which gives a realistic hope of realizing the 65 nm technology node by 2007, there have also been research demonstrations of smaller gate length devices. For example, conventional MOSFETs with gate lengths 15 nm [2.36], and 16 nm [2.37], which are required for hp45 nm technology node; 14 nm [2.38] for hp32 nm and 6 nm [2.39], 8 nm [2.40] and 10 nm [2.41] required for the hp22 nm technology node, have been fabricated and reported, delivering promising device parameters.

There is, however, growing consensus among the industry and research communities alike, that in the 45 nm technology node and beyond, it will become necessary to replace the conventional MOSFETs with novel device architectures, silicon on insulator (SOI), multiple gate FETs and wide application technology boosters such as strained silicon for carrier transport enhancement, high-κ gate stack, metal gates, etc.

2.3 Factors limiting the scaling of conventional MOSFET

Some of the technology advances enabled the scaling of conventional MOSFETs to decananometre dimensions were discussed in the previous sections. Unfortunately, even with all these advances the scaling of conventional MOSFETs whilst which maintaining good performance becomes increasingly difficult over time. Although the electronics industry has benefited from continuous scaling over the last three decades or so, the present trends indicate that the scaling of the conventional MOSFETs is fast approaching the end of its useful life time.

The optimistic prediction of the 2003 edition of the ITRS that scaling will continue until 2018 and beyond, is challenged by some fundamental limitations. Quantum mechanical effects such as carrier confinement and tunnelling, the

randomness of discrete doping, and worries over the increasing power dissipation are some of the main factors that may force the industry to a paradigm shift in MOSFET architecture and process technology. In the next section, some of the fundamental limitations to scaling are examined

2.3.1 Quantum mechanical tunnelling

The three main quantum mechanical tunnelling phenomena, which affect the MOSFET scaling, are illustrated in figure 2:9. They include band-to-band tunnelling, gate tunnelling, and source to drain tunnelling. All three tunnelling process are discussed below.

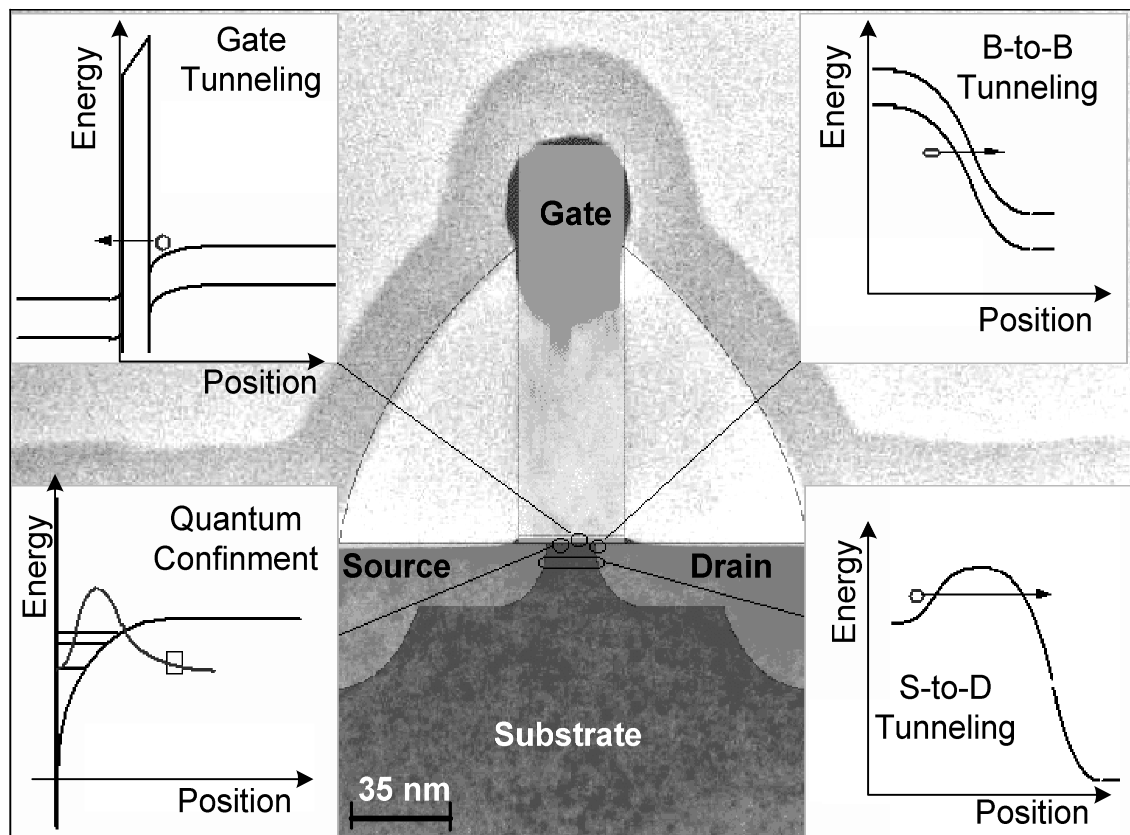

Figure 2:9 Visual illustrations of quantum effects near the Si/SiO2 interface: Reference [2.42]

2.3.1.1 Band to Band tunnelling

Band to band tunnelling (sometimes called Zener tunnelling) primarily occurs between the body and the drain of a MOSFET. The high channel doping that accompanies scaling results in a high electric field across the depletion layer at the reverse biased drain junction. The high electric field (~10^6V/cm) favours a parasitic leakage current associated mainly with the tunnelling of electrons from the valence band in the channel region to the conduction band in the drain [2.43] [2.8]. The tunnelling current density is approximated by:

$$J_{B-B} = \frac{\sqrt{2m^*}q^3 EV_{app}}{4\pi^3 \hbar E_g} \exp(-4\frac{\sqrt{2m^*}E_g^{3/2}}{3q\text{E}\hbar}) \quad (2.9)$$

$$E = \sqrt{\frac{2qN_a(V_{app} + \psi_{bi})}{\varepsilon_{si}}} \quad (2.10)$$

where m^* is the electron effective mass, E is the electric field, V_{app} is the applied reverse voltage across the junction, E_g is the energy gap, ψ_{bi} is built in potential and $\hbar$ is a modified Planks constant ($\hbar = h/2\pi$). Due to transistor scaling the increased doping concentration increases the electric field in equation (2:10), which also increases the tunnelling current in (2.9). This is depicted in figure 2:8 below.

The leakage current due to band to band tunnelling is less than the other two leakage currents (off state and gate leakage) measured at the physical gate length of 30nm. Nevertheless, it is clear from the overall tendency of the junction leakage current shown in figure 2.10, that for smaller gate lengths its contribution to the total leakage current will become significant.

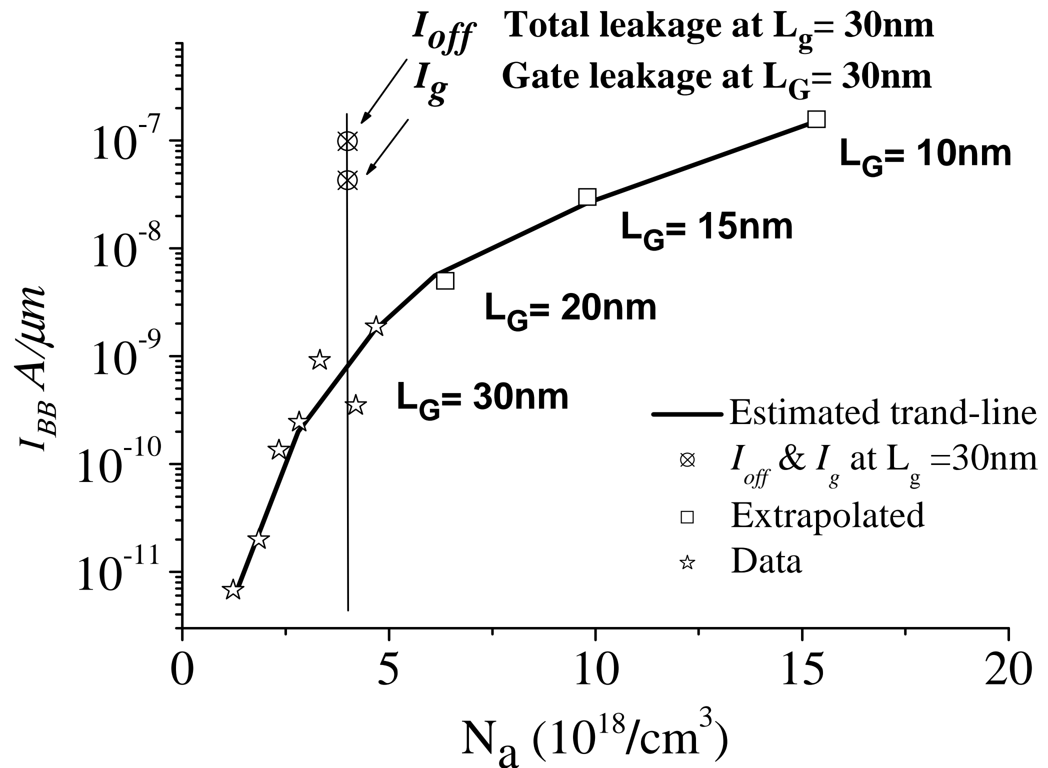

Figure 2:10 Junction leakage current due to band to band tunnelling (reference [2.44])

2.3.1.2 Direct gate oxide tunnelling

In order to achieve a desired current drive at a substantially low power supply voltages in sub 50 nm MOSFETs, aggressively scaled gate-dielectrics with equivalent oxide thickness (EOT) in the range of $t_{ox} \approx 1.5 - 0.5nm$ are required according to the latest edition of the ITRS. For such ultra-thin oxides the channel carriers can tunnel into the polysilicon gate through the gate-dielectric material. This process of electron or hole transmission through the dielectric barrier increases the gate leakage current exponentially with decreasing t_{ox} [2.45]. As a consequence of the increase in gate current, the overall off-state current is raised to an intolerable *l*evel for real circuit applications.

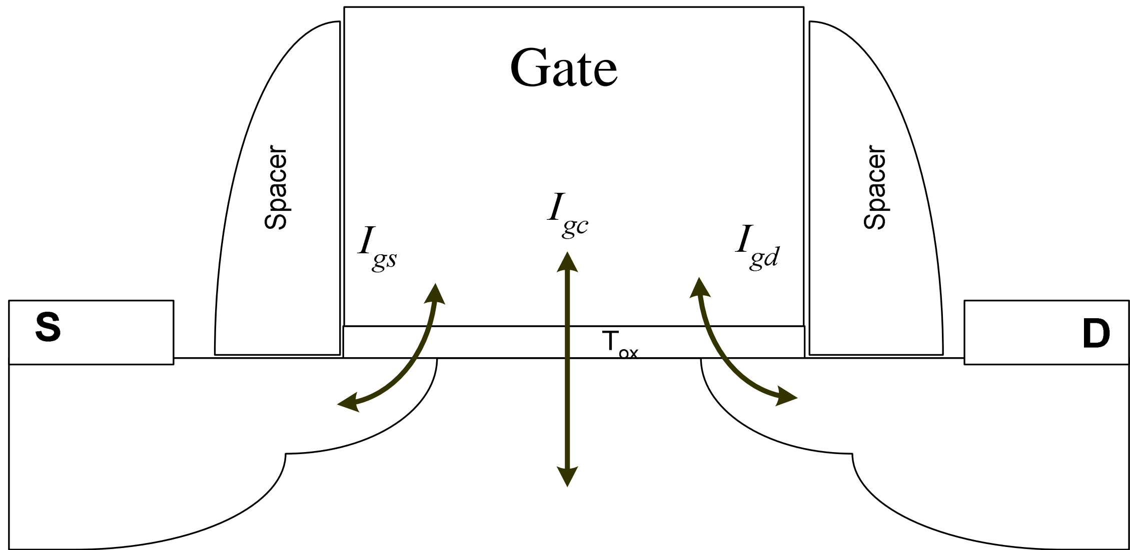

Figure 2:11 The overall gate tunnelling current is the sum all the components tunnelling current namely the source-to-gate, drain-to-gate and channel to gate currents (reference [2.46])

In the previous generations of MOSFET devices, the contribution of gate oxide leakage current to the overall leakage current has not been substantial. However, this is not the case for the present and the next generation of devices. As shown in figure 2.11, in addition to the tunnelling current from the channel I_{gc}, the fringing currents from the gate overlap with the source extension-I_{gs} and drain extension-I_{gd} also contribute to the total gate leakage current [2.46]. Bearing in mind that the main source of standby power drawn in CMOS circuits originates from the off-state current, substantial increase in off current due to direct tunnelling through the gate dielectric should be considered seriously during the design of devices or power optimization in circuits [2.47]. The application dependant power constraint as one of the main limitations of MOSFET scaling will be discussed in section 2.3.3

The cumulative gate tunnelling current density can be approximated by equation (2.4) [2.45]:

$$J_{DT} = \frac{4\pi q m_1 kT}{h^3} \int_0^{E_b} \tau_C(E) \ln\left[\frac{[\exp(E_{Fn1} - E_{C1} - E)/kT] + 1}{[\exp(E_{Fn2} - E_2 - E)/kT] + 1}\right] dE \qquad (2.11)$$

where m_1, and is the electron effective tunnelling mass E_{Fn1} the electron Fermi level, and E_{c1} is the bottom of conduction band in Si, and E_{Fn2}, and E_{c2} in the polysilicon region.

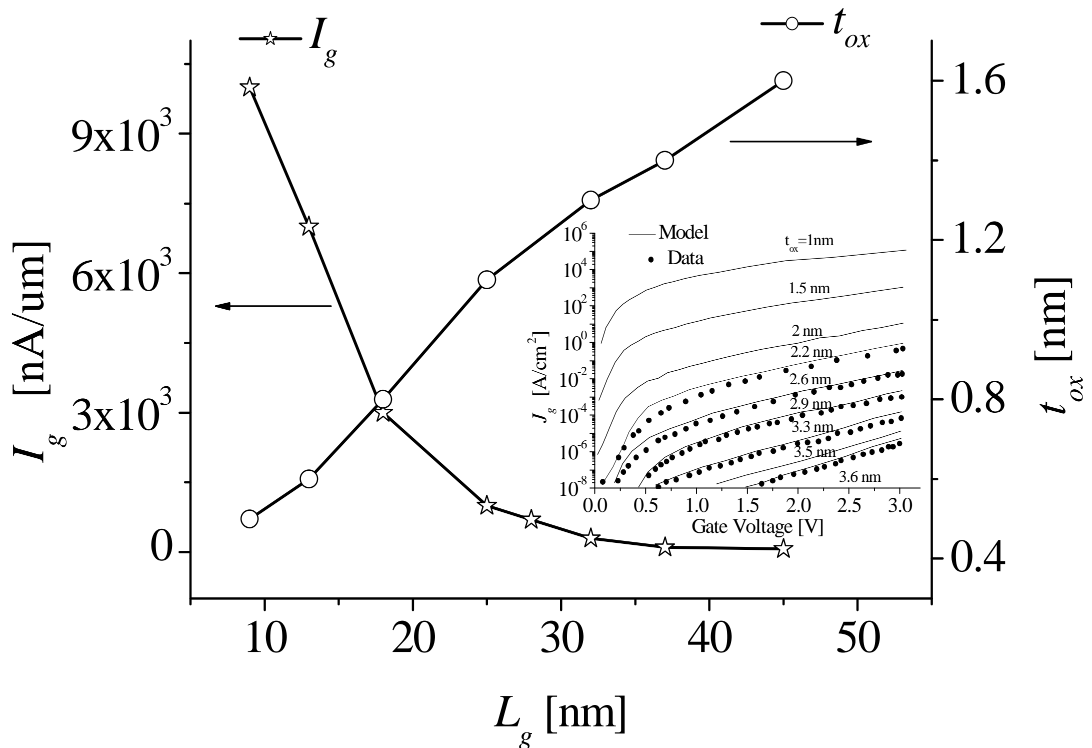

Figure 2:12 The relationship between the gate leakage current and the physical gate length. Inset: the gate current density for various oxide thicknesses as a function of gate voltage. Data from ITRS (2004 update) and the inset graph is adapted from reference. [2.48]

Figure 2:12 illustrates the relationship between gate length reduction and the gate oxide thickness which are required for the 90nm technology node and beyond. The figure also shows the increase of the gate leakage current density for various oxide thicknesses. The increase in leakage current is exponential [2.49]. For example, for gate lengths below 10 nm, and corresponding oxide thicknesses of 0.4 - 0.5, nm the leakage current reaches about 8-10 μA/μm provided that SiO_2 is used as a gate dielectric material. This amount of leakage current can affect transistor integration in high performance digital systems where low power dissipation per area is a critical issue during the stand-by mode of operation.

The inset in figure 2:12 compares according to [2:48] the experimental data and the calculated gate tunnelling current density for a range of oxide thickness as a function

of the gate voltage. It illustrates well the exponential increase in the gate tunnelling with the reduction of the oxide thickness for the whole range of relevant gate voltages.

2.3.1.3 Source to drain tunnelling

The other possible cause of tunnelling current that can affect the operation of sub-10 nm MOSFETs is source-to-drain tunnelling. The proximity of the source and drain metallic-junctions may lead to quantum mechanical tunnelling that will increase the overall transistor leakage current. The effect of source to drain tunnelling current in an 8 nm MOSFET has been experimentally demonstrated by Kawaura *et al.* [2.50]. According to the technology roadmap such devices will be in production around year 2018.

The comparison of the subthreshold current and its temperature dependence made in [2:50] for two experimental devices with $L_g = 52$ nm and $L_g = 8$ nm shows that there is an increase in subthreshold leakage current and degradation of the subthreshold slope in an 8 nm transistor due to source to drain tunnelling. Although at room temperature the effect is negligible, the contribution of the source to drain leakage current at the lower temperatures increases. Due to the relatively low temperature sensitivity of the tunnelling current at low temperature when it dominates the subthreshold current, the subthreshold slope becomes temperature independent. It has also been shown in [2.50] that when source to drain tunnelling starts to play a big role the sub threshold slope becomes significantly degraded. This is also reported in [2.51] based on a density gradient device simulation approach.

2.3.2 Intrinsic parameter fluctuations

In addition to the investigation of the scaling properties of MOSFETs, the investigation of intrinsic parameter fluctuations in scaled MOSFETs is also considered as part of this project. Statistical simulation results of the effect of different sources of intrinsic parameter fluctuations, together with detailed discussions, are presented in chapter five. In this section, a background discussion on intrinsic parameter fluctuations in decananometre MOSFETs, and a brief review of some previous results are presented.

According to the latest update of the ITRS'04, transistors with gate length of 7nm are required for the 18 nm technology generation in the year 2018. The channel length of these devices is approximately 10-14 silicon atoms in span. The random number and position of dopants from device to device in such a small nominal volume of the crystalline lattice will introduce significant variation in key device parameters (such as threshold voltage and off-current) in any ensemble of devices. Gate line edge roughness (LER) and the oxide thickness variations are two other sources of intrinsic parameter fluctuations [2:52][2:53]. It has been shown, that for devices comparable to in size to the 35 nm MOSFET, interface roughness introduces significantly less fluctuations compared to the those from doping or LER, and will not be discussed in this work.

The intrinsic parameter fluctuations associated with random dopants in CMOS have been widely investigated over the last two decades. Initially the effect of random discrete dopants on V_T was studied by Hoeneisen and Mad in [2:54], which showed that a non uniform distribution of random dopants in the channel causes mismatch in the V_T of CMOS device. Hagiwara *et al*., in the early 80's, investigated the effect of random dopants analytically and suggested a simple model to estimate V_T variation [2:55]. Furthermore, intrinsic parameter fluctuations in V_T, have been experimentally demonstrated and reported in [2.56] [2.57]. They have also been extensively studied theoretically based on statistical 3-D atomistic simulations by Asenov *et.al*., [2.58 2:59], by Wang and Taur [2.60] and using a 3-D Monte Carlo approach by Frank *et al*., [2:61].

Mizuno *et al*., [2.57] have shown experimentally that the V_T fluctuations have a Gaussian distribution, which is a typical characteristic of random events. They also confirmed experimentally that the threshold voltage fluctuations increase with reduction of the channel length. The magnitude of the V_T fluctuations in their experimental device with a channel length of $L_{eff} = 500nm$ was about $\pm 3\sigma V_T = \pm 12$mV, which accounts for 60% of the overall fluctuations. Although this value is smaller than present fluctuations magnitude, its contribution to the overall fluctuations in devices is significantly high.

Extrapolation of these fluctuations, based on simulation studies, to decanano-meter device dimensions shows a very significant increase. For example, in the 35nm channel length MOSFET, the $\pm 3\sigma V_T$ increase up to ± 100mV [2.62] and are significantly larger for smaller gate length devices, as will be shown in chapter 5.

2.3.3 Power dissipation

Aside from the physical scaling discussed previously in this chapter, application dependant power dissipation becomes one of the major factors hampering the integration of the scaled devices and therefore limiting the usefulness of continued scaling [2.63].

From a VLSI circuits' application point of view, there are three power dissipation mechanisms, dynamic or switching power, short current generated power, and static power dissipations. These added together, give the total power dissipation, P_T, on integrated circuit:

$$P_T = \sum (P_D, P_{sh}, P_{st}) \quad (2.15)$$

Where $P_D = CV_{dd}^2 f$ is the dynamic (active) power dissipation due to the charging and discharging of the capacitive load on each one of the devices in the integrated circuit. The dynamic power dissipation takes place during transition from high to low logic level or vice versa. The second term in equation (2.15) $P_{sh} = V_{dd} I_{short}$ is the power dissipation due to occurrence of short circuit currents. This depends generally on the architecture of the circuit. Usually this is not the main component of power dissipation. The third component of the total power dissipation is the static power dissipation, $P_{st} = V_{dd} I_{leak}$, which occurs as a result of the cumulative leakage current contributed from all devices in the circuit.

By far the highest power consumption in present circuits comes from the dynamic activity of the devices in the circuit. The dynamic power dissipation is directly proportional to the square of the supply voltage. Although the power supply voltage is reducing as a result of device scaling, due to this quadratic relationship between P_D and V_{dd}, dynamic activity, still accounts for substantial power dissipation. However, the static power dissipation is gathering pace as the major power constraint in deeply scaled MOSFETs. This is illustrated in figure 2:13 which shows that the proportion of the static power dissipation in the 65 nm technology node circuits is increased compared to previous technology nodes.

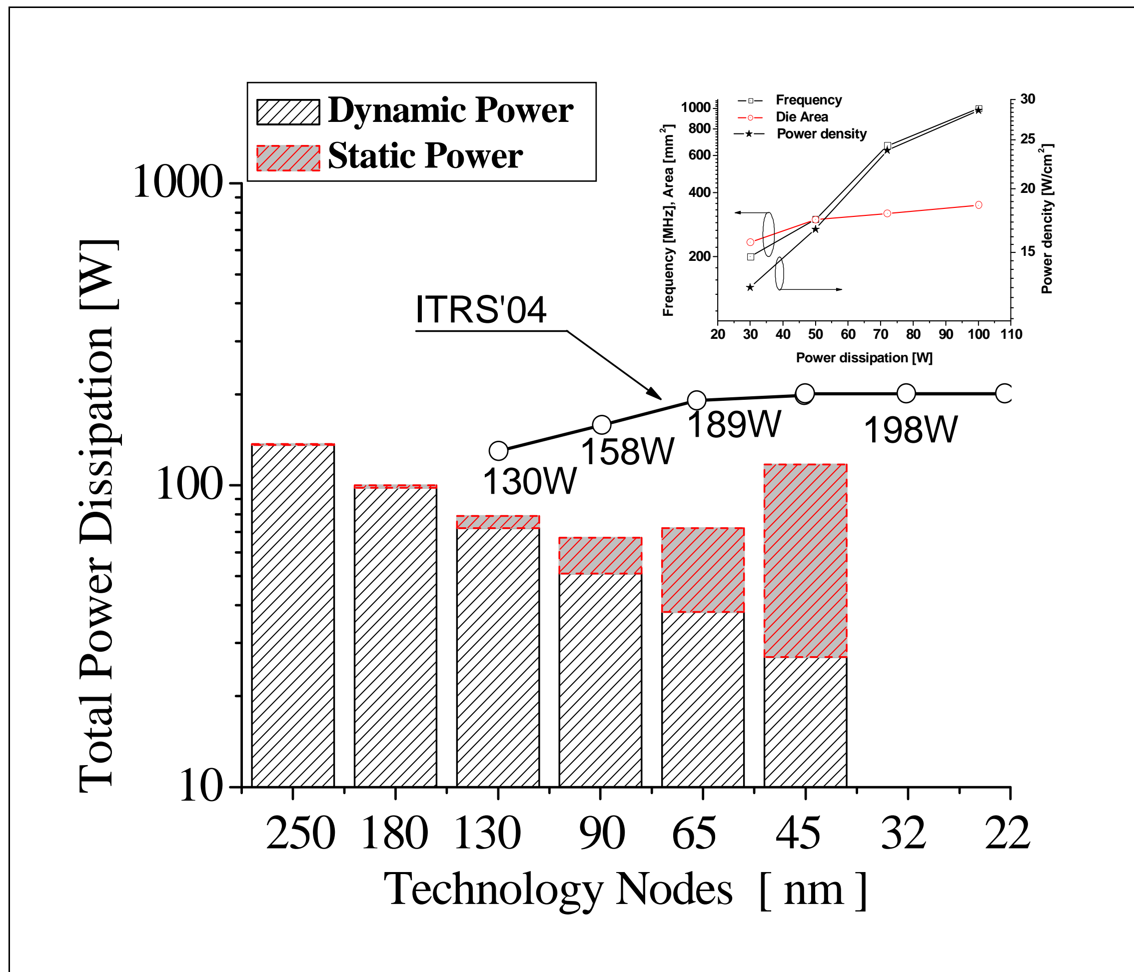

Figure 2:13 Total power dissipation as a function of technology node and an illustration of the increasing of the static power dissipation due to the significant increase in the off-state current as technology scales to 65nm and beyond. Data source: ITRS 2004 and references [2.64] [2.65]

For the 45 nm technology node the static power dissipation is expected to be more than 50% of the total power consumption (see figure 2:13). The ITRS requirement of maximum allowable power dissipation is also indicated in the graph and has been kept constant (P_T=198W) beyond the 65 nm technology node. The experimental data shown in the bar graph corresponds to 30 metres wide transistors for each technology generation [2.64].[**]

The inset in figure 2:13 depicts the data reported in [2.65] on the relationship between the power dissipation and the frequency. It also shows how the power density

[**] This is equivalent to 30 million transistors by considering the width of a single transistor is 1µm.

P_T / A is increasing significantly as the number of components integrated into a chip increases.

As has been discussed in section 2.3.1, the main scaling limiting factors are the different component of the leakage currents. For high performance devices near the end of the ITRS all three major tunnelling leakage currents (gate oxide, source-to-drain, and band-to-band tunnelling of electrons or holes) will contribute to the ever increasing power density of VLSI circuits. The reason is that the power supply voltage for high performance devices is relatively higher than for the low operating power (LOP) transistors. Therefore, controlling the leakage current in individual devices and design of power optimized VLSI circuit architectures remain the main issue for research and development stages of the next generation devices.

2.3.4 Reliability of the Ultrathin Gate Oxide

Reliability engineering is an important aspect of CMOS technology. It requires meticulous measurement, rigorous testing and careful statistical data analysis of the performance of the MOSFETs and the other circuit components. The outcome of this process enables the engineers to quantify the reliability of a given integrated circuit.

The reliability of SiO_2, which is still the main gate dielectric material used for the fabrication of CMOS devices in the semiconductor industry, has been extensively studied over the past three decades. Most of the research has been concentrated mainly on trap creation (defect generation) mechanisms that degrade the gate oxide and eventually lead to its breakdown [2:66, 2:67, 2:68, 2:69]. For the sake of completeness, and to highlight in particular the reliability issues connected with possible breakdown of the aggressively scaled ultra-thin ($t_{ox} \le 2nm$) gate dielectric materials, a brief discussion is presented in this sub-section.

Reliability in general is associated with the duration of time over which the device is operating at its full potential, or within some degree of tolerance. This period is called the "time-to-break down" t_{BD} which is commonly known as the product life-time. With regard to the gate dielectric material reliability, it can also be associated with the charge build-up in time (related to the "charge-to-break down", $t_{BD} = Q_{BD}/J$, where J is gate leakage current) and Q_{BD} is given by equation (2.8) [2.70]

$$Q_{BD} = \frac{N_{BD}}{P_{gen}} \tag{2.14}$$

Where N_{BD} is the defect density which trigger breakdown and P_{gen} is the defect generation rate at which the defects are created and is given by (2.15)

$$P_{gen} = \left(\frac{\Delta J}{J}\right)\frac{1}{Q_{inj}} \tag{2.15}$$

Here $\Delta J/J_0$ is the relative change of the stress induced leakage current and Q_{inj} is the injected electron charge.

The main focus of this section is to relate the oxide thickness to reliability. However, equations 2.14 and 2.15 are not explicit function of the oxide thickness (t_{ox}). DiMaria *et al*, [2.71] have found that the average density of states to breakdown (N_s^{BD}) *decreases* with decreasing t_{ox}. This means that fewer microscopic defects are enough to cause critical breakdown in very thin gate dielectric materials Furthermore, Degraeve *et al*. [2:72], have investigated the dielectric material thickness dependence of the Weibull slope (β) defined below. They show that the Weibull slope becomes shallower with decreasing t_{ox}. The gate oxide thickness dependence of the Weibull slope is also reported in [2.73] using the cell-based statistical breakdown model. The Weibull function describes the cumulative probability of failure *F,* which is given by 2.16 [2.74]:

$$F(Q_{BD}) = 1 - \exp(\frac{Q_{BD}}{\alpha})^{\beta} \tag{2.16}$$

Equation (2.16) is often rewritten as in the form

$$\ln\left(\ln\left(1 - F(Q_{BD})\right)\right) = \beta \ln(\frac{Q_{BD}}{\alpha}) \tag{2.17}$$

where β is known as the shape factor of the distribution (often referred to as the Weibull slope) and α is the characteristic charge (or time) to breakdown at the reliability rate of 37% (or 63% failure percentage).

Figure 2:14 illustrates the relationships between the critical defect density (N_{BD}) and the oxide thickness. The critical defect density is strongly dependant on the oxide thickness and the amount of the charge to breakdown is highly dependant on the stress voltage.

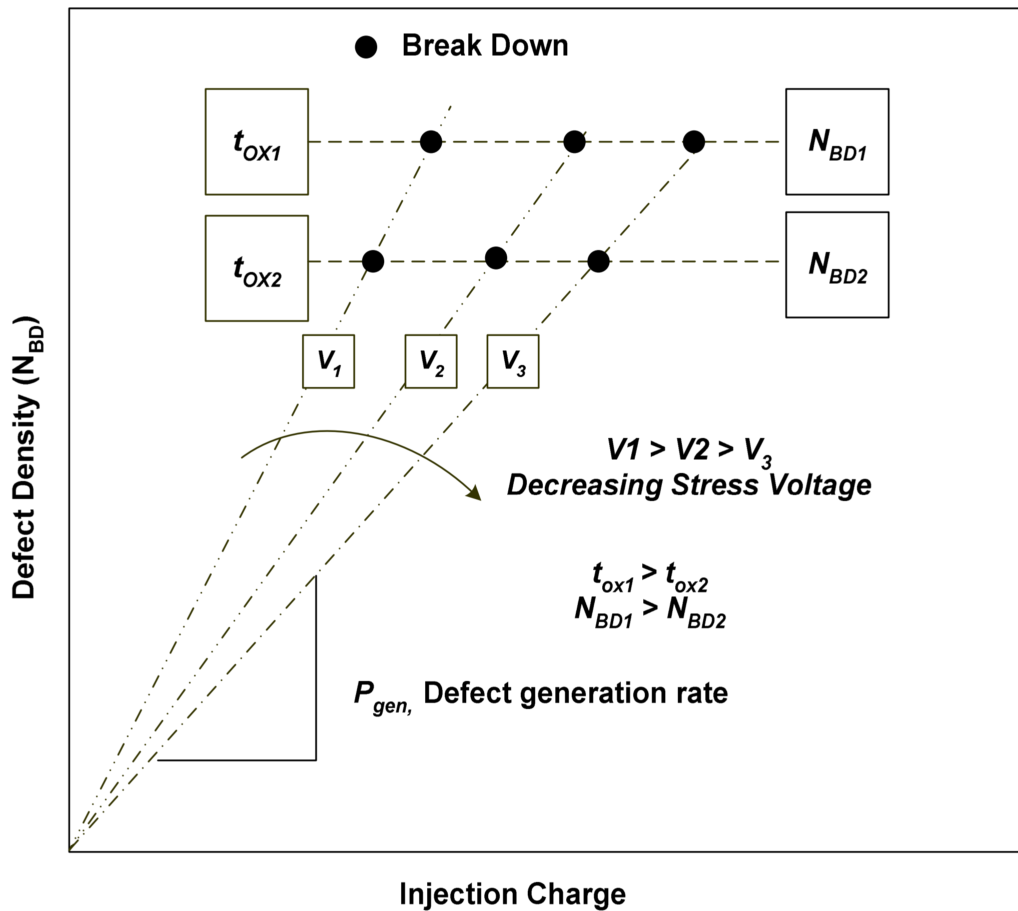

Figure 2:14 Schematic illustration of the relationship between the defect generation rate, defect density, oxide thickness and the stress voltage: Reference [2.70]

2.4 Chapter Summary

In this chapter we have reviewed Moore's law and the International Technology Roadmap for Semiconductor Technology. Both Moore's law and the ITRS have played a significant role in the advancement of the CMOS technology. In particular the road map has been an important guiding document that gives the direction for research and development for next generation devices. Apart from their role as a driving force behind the development of semiconductor technology as a whole, our emphasis was on their influence on MOSFETs scaling

The two most important scaling rules, 'constant field' and the 'generalised scaling,' have been discussed in detail. Based on the physical properties of short channel MOSFETs, the generalised scaling rule has been adopted for this work. It takes into account the 2D properties of the electric field and the retarded reduction of the supply voltage during the scaling process.

The limiting factors that hinder the progress of MOSFET scaling have also been discussed. Application dependant power constraints (including the leakage currents), the intrinsic parameter fluctuations, and the issue of the reliability of ultra-thin gate dielectric material are among the major scaling limiting factors that jeopardise Moore's law in the near future.

The main ultimate limitation to the scaling of transistors is the atomic structure of matter. A MOSFET cannot be smaller than an atom. The transistor post 2025, according to the IBM roadmap, will be reaching these limitations.

In the next chapter the simulation strategy and overall methodology used to investigate the scaling and intrinsic parameter fluctuation in nano CMOS devices are presented.

2.5 Chapter reference

[2.1] G. E. Moore, "Cramming more components onto integrated circuits", *Electronics*, Vol. 38, number 8, April 19, 1965

[2.2] J. Meindl, "Theoretical Practical and Analogical Limits in ULSI", *IEDM Tech. Digest*, IEEE press, pp. 8-13, 1983

[2.3] G. M. Borsuk, and T. Coffey, "Moore's Law: A Department of Defence Perspective", *Defence Horizons*, Number 30, July 2003

[2.4] International Technology Roadmap for Semiconductors, (SIA) http://public.itrs.net, SEMATECH, 2003 edition of ITRS

[2.5] G. E. Moore, "Progress in Digital Integrated Electronics", *IEDM Tech. Digest*, *IEEE press*, pp. 11-13, 1975

[2.6] R. Nair, "Effect of increasing chip density on the evolution of computer architecture", *IBM J. Res. & Dev.*, Vol. 46, No. 2/3, pp. 223-234, March/May 2002

[2.7] P. P. Gelsinger, P. A. Gargini, G. H. Parker, A. Y. C. Yu, "Microprocessor circa 2000", *IEEE SPECTRUM*, pp. 43-47, October 1989

[2.8] Y. Taur, and T. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, United Kingdom, 1998

[2.9] E.J Novak, "Maintaining the benefits of CMOS scaling when scaling bogs down", *IBM J.Res & Dev.*, Vol. 46, NO. 2/3, 2002

[2.10] G. Baccarani, M.R. Wordeman, and R.H. Dennard, "Generalised Scaling Theory and its Application to ¼ Micrometer MOSFET design," *IEEE Trans., Electron Devices*, Vol.31, pp. 452, 1984

[2.11] J. R. Brews, *et. al*. "Generalised Guide for MOSFET Miniaturization, " *IEEE Electron Device letter*, Vol. EDL-1, pp. 2-4, 1980

[2.12] S. Inab, K. Okano, S. Matsuda, M. Fujiwara, A. Hokazono, K. Adachi, K. Ohuchi, H. Suto, H. Fukui, T. Shimizu, S. Mori, H. Oguma, A. Murakoshi, T. Itani, T. Iinuma, T. Kudo, H. Shibata, S. Taniguchi, M. Takayanagi, A. H. Oyamatsu, K. Suguro, .Y. Katsumata Y. Toyoshima, and H. Ishiuchi, "High Performance 35 nm Gate Length CMOS With NO Oxynitride Gate Dielectric

and Ni Salicide", *IEEE Trans. Electron Devices*, Vol. 49(12), pp. 2263 – 2270, 2002

[2.13] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H-S. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies", *Proceedings of The IEEE*, Vol. 89, No. 3, pp. 259-288, March 2001

[2.14] R.H. Dennard, F. H. Gaensslen, L. Khun and, H. N. Yu, "Design of micron switching devices", *Presented at IEDM*, Washington D.C., December 1972

[2.15] R.H. Dennard, F. H. Gaensslen, Hw A-Nien Yu, V. L. Rideout, E. Bassous, and A. R, LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions", *IEEE Journal of Solid-State Circuits*, Vol. SC-9, No. 5, pp. 256-268, 1974

[2.16] N. D. Arora, J. R. Hauser, and D. J. Roulston, "Electron and Hole Mobilities in Silicon as a Function of Concentration and Temperature," *IEEE Trans Electron Devices*, Vol. Vol. 29, pp. 292-295, Feb. 1982.

[2.17] K. K. Ng, S. A. Eshraghi, and T. D. Stanik, "An Improved Generalised Guide for MOSFET Scaling", *IEEE Trans Electron Devices*, Vol. 40(10), pp. 1895-1897, 1993

[2.18] T. Skotnicki, "Heading for decanano CMOS- Is navigation among the iceberg still a viable strategy?", *Proceedings of ESSDERC2000*, pp. 19-33, Sep. 2000

[2.19] T.Skotnicki, C. Denat, P. Senn, G. Merckel, and B. Hennion, "A new analog/digital CAD model for sub-halfmicron MOSFETs", *IEDM Tech. Dig*., pp. 165 - 168, December 1994

[2.20] D. J. Frank, "Power-constrained CMOS scaling limits", *IBM J. Res. & Dev.,* Vol. 46, No. 2/3, pp. 235-244, March/May 2002,

[2.21] C-T. Sah, "Evolution of the MOS Transistor –From Conception to VLSI", *Proceedings of the IEEE*, Vol. 76 (10), pp. 1280-1326, 1988

[2.22] J. R. Burns, "Switching response of complimentary MOS transistor logic circuit", *RCA Review*, Vol. 25, pp. 627-661, December 1964

[2.23] A.K. Rapp, L. P. Wennick, H. Borkan and K. R. Keller, "Complementary-MOS Integrated binary Counter", *1967 International solid-state circuit conference, Digest of technical papers*, pp.52-53, 1967

[2.24] T. D. Kamins, "Preparation and Polycrystalline-Silicon Films ",in Handbook of Semiconductor Silicon Technology, Edited by W. C. O'Mara, Robert B. Herring, and L. P. Hunt, Noyes Publications, pp. 640-730, 1990

[2.25] K. Saito, T. Morose, S. Sato, and U. Harada, "A new short channel MOSFET with lightly doped drain", *Danshi Tsushin Rengo Taikai*, pp. 220, April 1979

[2.26] S. O., Paul, J. Tsang, W. W. Walker, L. D. Critchlow, and J. F. Shepard, "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor", *IEEE J. Solid-State Circuit*, Vol.SC-15(4), pp. 424-432, 1980

[2.27] P. J. Tsang, S. Ogura, W. W.Walker, J. F. Shepard and L. D. Critchlow, "Fabrication of High-Performance LDDFET's with Oxide Sidewell-Spacer Technology ", *IEEE J. Solid-State Circuit*, Vol.SC-17(2), pp. 220-432226, 1982

[2.28] K.-H. Oh, Z. Yu, and R. W. Dutton, "A Bais Dependant Source/Drain Resistance Model in LDD MOSFET Devices for Distortion Analysis", *6th International Conference on VLSI and CAD: ICVC'99*, pp. 190-193, 1999

[2.29] B. L. Crowder and S. Zirinsky, "1-micron MOSFET VLSI technology. VII - Metal silicide interconnection technology: A future perspective", *IEEE J. Solid-state Circuits,* SC-14, pp. 291-293, 1979

[2.30] G. G. Shahidi, D. A. Antoniadis, and H. I. Smith, "Indium channel implant for improved short-channel behaviour of sub micrometer NMOSFET's," *IEEE Trans. Electron Devices*, vol. 14, p. 409, Aug. 1993.

[2.31] I. De and C. M. Osburn, "Impact of Super-Steep-Retrograde Channel Doping Profiles on the Performance of Scaled Devices", *IEEE Trans. Electron Devices*, vol. 46, p. 1711-1717, 1999

[2.32] Y. Taur, C. H. Wann, and D. J. Frank, "25nm CMOS Design Consideration", IEDM'98 Technical Digest, pp. 789-792, 1998

[5:33] S. Thompson, *et al.*, "A 90nm technology fearturing 50nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD and 1um 2SRAM cell ," 2002 IEDM Digest, pp. 21-64, Dec. 2002

[2:34] W. M. Siu, "Forward", *Intel Tech. J.* Vol 8 (1), 2004

[2:35] K. Goto, *et al.*, *IEDM Tech Dig*, IEDM'03, pp. 623, 2003

[2.36] B. Yu, H. Wang, A. Joshi, Q. Xiang, E. Ibok, M.-R. Lin, "15nm Gate length Planer CMOS Transistor", *IEDM Tech Dig*, IEDM'01, pp. 937 - 939, 2001

[2.37] F. Boeuf , T. Skotnicki , S. Monfray , C. Julien 1, D. Dutartre, J. Martins , P. Mazoyer1, R. Palla, B.Tavel, P. Ribot, E. Sondergard, and M. Sanquer, "16nm planar NMOSFET manufacturable within state-of-the-art CMOS process thanks to specific design and optimisation", *IEDM Tech Dig*, IEDM'01, pp. 637 - 640, 2001

[2.38] A. Hokazono *et al.*, "14 nm Gate Length CMOSFETs Utilizing Low Thermal Budget Process with Poly-SiGe and Ni Salicide", *IEDM Tech Dig*, IEDM'02, pp. 639 - 642, 2002

[2.39] N. Yasutake, *et al.*, "A hp22nm Node Low Operating Power (LOP) Technology with Sub-10nm Gate lLength Planar Bulk CMOS Devices", Digest of Technical Papers, 2004 Technology Symposium, pp. 84, 2004

[2.40] B. Doris, *et al*, "Device Design Considerations for Ultra-Thin SOI MOSFETs", *IEDM Tech Dig*, IEDM'03, pp. 631 - 634, 2003

[2.41] H Wakabayashi, "Transport Properties of Sub-10-nm Planar-Bulk-CMOS Devices", *IEDM Tech Dig*, IEDM'04, pp. 429 - 432, 2004

[2.42] A. Asenov, J. R. Watling A. R. Brown, D. K. Ferry, "The use of Quantum Potential for Confinement and Tunnelling in Semiconductor devices", *Journal of Computational Electronics,* 1:503-513 2002

[2.43] S. M. Sze, Semiconductor devices : Physics and technology, John Willey & Sons, INC, 2nd Edition, 2001

[2.44] B. Doyle *et. al.*, "Transistor Elements for 30nm Physical Gate Length and beyond," Intel *Technical Journal*, Vol. 06, Issue 02, pp. 42, 2002

[2.45] S. –Lo, D. Buchannen, Y. Taur, and W. Wang, "Quantum-Mechanical modelling of Electron Tunnelling Current from Inversion Layer of Ultra-Thin-Oxide nMOSFETs", *IEEE Electron Device Lett.*, Vol. 18, pp. 209-211, 1997

[2.46] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. W. Dutton, "Impact of Gate Direct Tunneling Current on Circuit Performance: A Simulation Study", *IEEE Trans, Electron Devices*, Vol. 48(12), pp. 2823-2829, 2001

[2.47] D. Lee, D. Blaauw, and D. Sylvester, "Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits" *IEEE Trans. Very Large Scale Integration(VLSI) Systems*, Vol. 12(2), pp. 155-166, 2004

[2.48] Y. Taur, "CMOS design near the scaling limit" *IBM J. Res. Develop*, Vol. 46(2/3), pp. 213-222, 2002

[2.49] S. –H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of quantization, polysilicon depletion, and direct tunnelling effects in MOSFETs with ultra thin oxides", *IBM J. Res. Develop*, Vol. 43(3), pp. 327-337, 1999

[2.50] H. Kawaura, T. Sakamoto, and T. Baba, "Observation of source-to-drain direct tunneling current in 8 nm gate electrically variable shallow junction metal–oxide–semiconductor field-effect transistors", *Applied Physics Letter*, Vol. 76, NO. 25, pp. 3810 - 3812, 2000

[2.51] J. R. Watling, A. R. Brown, and A. Asenov, "Can the Density Gradient Approach Describe the Source-Drain Tunneling in Decanano Double-Gate MOSFETs", *Journal of Computational Electronics*, 1:289-293 2002

[2:52] A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic Threshold Voltage fluctuations in Decanano MOSFETs Due to Local Oxide Thickness Variations", *IEEE Tran. Electron Devices,* Vol. 49 (1), pp. 112-119, 2002

[2.53] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic Parameter Fluctuations in Decananometer MOSFETs Introduced by Gate Line Edge Roughness", *IEEE Tran. Electron Devices,* Vol. 50(5), pp. 1254-1260, 2003

[2:54] B. Hoeneisen and C. A. Mad, "Fundamental Limitation in Microelectronics –I. MOS Technology", *Solid-State Electronics*, Vol. 15, pp. 819-829, 1972

[2:55] T. Hagiwara, K. Yamaguchi, and S. Asai, "Threshold voltage variation in very small MOS transistors due to local dopant fluctuations," *Proc. Symp. VLSI Technol., Dig. Tech. Papers,* 1982, pp. 46–47.

[2.56] R.Kadaba, R. Lakshmikumar, A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design", *IEEE journal Of Solid-State circuits,* Vol. Sc-21,No. 6, pp. 1057, 1986

[2.57] T. Mizuno, J –I. Okamura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuation Due to Statistical Variation of Channel Dopant Dopant Number in MOSFET's", *IEEE Trans, Electron Devices*, Vol. 41(11), pp. 2216-2221, 1994

[2.58] A. Asenov, "Statistically reliable 'Atomistic' simulation of Sub 100nm MOSFETs", in Simulation of Semiconductor Process and Devices, edited by K. De Meyer and S. Biesemans, Springer, Wien New York, pp. 223-226, 1998

[2:59] Asen Asenov, Andrew R. Brown, John H. Davies, Savas Kaya, and Gabriela Slavcheva, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs", *IEEE Trans, Electron Devices*, Vol. 50(9), pp. 1837-1852, 2003

[2.60] H. S. Wong, and Y. Taur, "Three dimensional 'atomistic' simulation of discrete random dopant effects in sub-0.1 mm MOSFETs", *Proc. IEDM'93 dig. Tech. Papers*, pp. 705, 1993

[2:61] D. J. Frank, Y. Taur, M. Ieong, and H-S P. Wang, "Monte Carlo Modeling of Threshold Variation due to Dopant Fluctuation", *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 171-172, 1999

[2.62] F. adamu-Lema , G. Roy, A. R. Brown, A. Asenov, and S. Roy, "Intrinsic Parameter Fluctuations in Conventional MOSFET's at the Scaling Limit: A Statistical Study", *Abstracts to IWCE10*, pp.44-45, 2004

[2.63] D. J. Frank, "Power constraint CMOS scaling limits", *IBM J. RES. DEV*. Vol. 46 (2/3), pp. 235-244, 2002

[2.64] B. Doyle *et al*, "Transistor Elements for 30nm Physical Gate Lengths and Beyond", *Intel Technology Journal*, Vol. 6(2), pp. 42-54, 2002

[2.65] T. Mudge, "Power: A first-Class Architect Design Constraint ", *IEEE Computer*, Vol. 34, pp 52-58, April 2001

[2.66] A. Berman, "Time Zero Dielectric reliability Test by a Ramp Method"

[2.67] J. M. Aitken, and D. R. Young, "Electron trapping by radiation-induced charges in MOS devices", *J. Appl. Physics*, Vol. 47, pp. 1196-1198, 1976

[2.68] D. J. DiMaria and J. W. Stasiak, "Trap creation in silicon dioxide produced by hot electrons", *J. Appl. Physics*, Vol 55(6), pp. 2342-2356, March 1989

[2.69] R. Degraeve, "Oxide Reliability Issues", in *High Dielectric Constant Materials, VLSI MOSFET Application, Springer*, pp. 91-120, 2005

[2.70] D. A. Buchnaan, "Scaling gate dielectric: Materials, integration, and reliability", *IBM J. Res. Develop*, Vol. 43(3), pp. 245-264, 1999

[2.71] D. J. DiMaria and J. H. Stathis, "Explanation for the oxide thickness dependence of breakdown characteristics metal-oxide- semiconductor structures", *Appl. Physics Lett.* , Vol. 70(20), pp. 2708-2710, May 1997

[2:72] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, and H. E. Maes, "A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides", *IEDM Tech. Dig.*, pp. 863 - 866, 1995

[2:73] D. J. DiMaria, and J. H. Stathis, "Explanation for the oxide thickness dependence of breakdown characteristics of metal–oxide–semiconductor structures", *Appl. Phys. Lett.*, Vol. 70, pp. 2708–10, 1997

[2:74] E. Y. Wu. And R. -P. Vollertsen, "On the Weibull shape Factor of Intrinsic Breakdown of Dielectric Films and Its Accurate Experimental Determination-Part I: Theory, Methodology, Experimental Techniques", *IEEE Trans. Electron Devices*, Vol. 49(12), 2002.