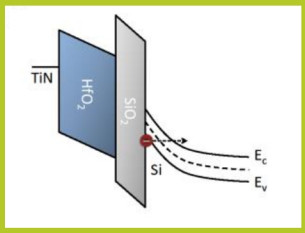
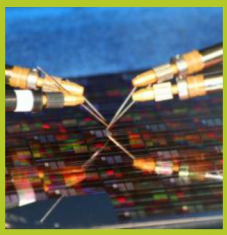
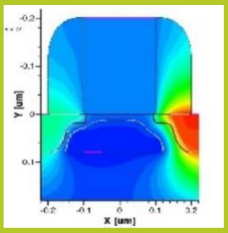
Research at

namlab



Reliability of high-k / metal gate field-effect transistors considering circuit operational constraints

Steve Kupke

Technische Universität Dresden

Reliability of high-k / metal gate field-effect transistors considering circuit operational constraints

Zuverlässigkeit von High-k Metal-Gate Feldeffekttransistoren unter Berücksichtigung von Schaltungsvorgängen

Steve Kupke

Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieurs

(Dr.-Ing)

genehmigte Dissertation

Vorsitzender:	Prof. Dr.-Ing. Henning Heuer	Tag der Einreichung: 01.09.2014
Gutachter:	Prof. Dr.-Ing. Thomas Mikolajick	Tag der Verteidigung: 19.03.2015
	Prof. Dr. rer. nat. Udo Schwalke	

"'It has today occurred to me that an amplifier using semiconductors rather than vacuum is in principle possible."'

William B. Shockley, laboratory notebook, 1939

Abstract

After many decades, the scaling of silicon dioxide based field-effect transistors has reached insurmountable physical limits due unintentional high gate leakage currents for gate oxide thicknesses below 2 nm. The introduction of high-k metal gate stacks guaranteed the trend towards smaller transistor dimensions. The implementation of HfO_2, as high-k dielectric, also lead to a substantial number of manufacturing and reliability challenges. The deterioration of the gate oxide properties under thermal and electric stress jeopardizes the circuit operation and hence needs to be comprehensively understood.

As a starting point, 6T static random access memory cells were used to identify the different single device operating conditions. The strongest deterioration of the gate stack was found for nMOS devices under positive bias temperature instability (PBTI) stress, resulting in a severe threshold voltage shift and increased gate leakage current. A detail investigation of physical origin and temperature and voltage dependency was done. The reliability issues were caused by the electron trapping into already existing HfO_2 oxygen vacancies. The oxygen vacancies reside in different charge states depending on applied stress voltages. This in return also resulted in a strong threshold voltage and gate current relaxation after stress was cut off.

The reliability assessment using constant voltage stress does not reflect realistic circuit operation which can result in a changed degradation behaviour. Therefore, the constant voltage stress measurement were extended by considering CMOS operational constraints, where it was found that the supply voltage frequently switches between the gate and drain terminal. The additional drain (off-state) bias lead to an increased V_t relaxation in comparison to zero bias voltage. The off-state influence strongly depended on the gate length and became significant for short channel devices.

The influence of the off-state bias on the dielectric breakdown was studied and compared to the standard assessment methods. Different wear-out mechanisms for drain-only and alternating gate and drain stress were verified. Under drain-only stress, the dielectric breakdown was caused by hot carrier degradation. The lifetime was correlated with the device length and amount of subthreshold leakage.

The gate oxide breakdown under alternating gate and off-state stress was caused by the continuous trapping and detrapping behaviour of high-k metal gate devices. This resulted in a lower lifetime in comparison to the standard DC and AC gate-only stress. The breakdown was primarily located close to the drain edge, expressed in a modification of the end of lifetime projection.

Supplementary, process treatments can positively influence the reliability. Fluorine is assumed to passivate oxygen vacancy and reduce the defect density. Fluorine was introduced at different positions within the gate stack by chemical vapor deposition. The Fluorine treatment revealed a reduction of the gate leakage current as well as threshold voltage shift being a promising candidate to improve the overall reliability.

Kurzfassung

Die fortwährende Skalierung von Polysilizium/Siliziumdioxid basierten Feldeffekttransistoren zur Steigerung der Schaltgeschwindigkeiten, hat ihre physikalischen Grenzen erreicht. Die geringe Gateoxiddicke (d < 2 nm) führt zu steigenden Gateleckströmen durch quantenmechanische Tunnelprozesse. Durch die Einführung von Hafniumdioxid als High-k-Dielektrikum mit einer metallischen Elektrode, konnte der Gateleckstrom, bei gleichbleibender Kapazität, stark reduziert werden. Die Einführung der High-k+Metal-Gate-Technologie (HKMG) führte zu einer Reihe von Veränderungen in der Transistorzuverlässigkeit, die Gegenstand der Untersuchungen in dieser Arbeit sind.

Anhand von Static Random Access Memory (SRAM) Transistoren wurden die verschiedenen Betriebszustände für n/pMOS Transistoren identifiziert und der elektrische Stressfall mit der höchsten Degradationsrate bestimmt. Im Gegensatz zum elektrisch negativen Gatestress (NBTI) bei Siliziumdioxid basierten pMOS Transistoren, zeigt der elektrisch positive Gatestress (PBTI) an HKMG n-Kanaltransistoren die grösste Veränderung der Schwellspannung und des Gateleckstroms. Das Degradationsverhalten, welches durch die Veränderung der Ladungszustände von Sauerstofffehlstellen im Hafniumdioxid hervorgerufen wird, wurde spannungs- und temperaturabhängig charakterisiert. Nach Spannungsstress, konnte eine Relaxation der Schwellspannungsänderung und Leckstromerhöhung beobachtet werden.

Bei der standardmässigen Zuverlässigkeitscharakterisierung unter Konstantspannungsstress kann der übliche dynamische Betrieb von Transistoren nicht berücksichtigt werden. Die Messmethoden wurden daher erweitert und das Degradationverhalten von Einzeltransistoren unter Schaltungsvorgängen (wechselndem Gate- und Drainstress) berücksichtigt. Diese zusätzliche Drainspannung beschleunigt die Relaxation nach Stress, wobei der Effekt stark von der Kanallänge abhängt und für Kurzkanaltransistoren relevant wird.

Dabei ergaben sich unterschiedliche Degradationsursachen für reinen Drainstress und Stress mit alternierender Gate- und Drainspannung. Der dielektrische Durchbruch unter hohen Drainspannungen wird durch Injektion von heissen Ladungsträgern hervorgerufen und ist stark kanallängenabhängig. Realistische Durchbruchsmessungen mit alternierenden Spannungen am Gate- und Drainkontakt führen zu einer geringeren Oxidlebensdauer im Vergleich zu normalen Gleich- und Wechselspannungmessungen am Gatekontakt. Die asymmetrische Feldverteilung führt zu einem dielektrischen Gateoxiddurchbruch an der Drainseite, was eine Anpassung der Modelle für die Lebensdauerbestimmung zu Folge hatte.

Zusätzliche Prozessierungsverfahren können einen positiven Effekt auf die Transistorzuverlässigkeit, wobei angenommen wird das Fluor zur Passivierung von Sauerstofffehlstellen beitragen kann. Mithilfe der chemischen Gasphasenabscheidung wurde Fluor an unterschiedlichen Positionen im Gatestapel eingebracht und der Einfluss auf die Defekteigenschaften von Hafniumdioxid untersucht. Die Fluorpassivierung resultierte in einem geringeren Gateleckstrom gegenüber der Referenzprobe, sowie in einem geringeren Degradationsverhalten der Schwellspannung unter Stress und stellt einen vielversprechenden Kandidat für die Verbesserung der Zuverlässigkeit dar.

Contents

1. Motivation

Theoretical principles are often ahead of experimental proof, which also holds true for the history of the transistor. In 1925, J. E. Lilienfeld described the first principle of a field-effect transistor in a patent [1]. It was not until 1947 that the first germanium point contact transistor came to realization by J. Bardeen, W. H. Brattain and W. S. Shockley [2], who were awarded for the Nobel Prize nine years later. Shockley continued to develop the bipolar junction transistor (BJT) [3], which became the most commonly used transistor type over three decades. In the late 1950's, Jack Kilby and Robert Noyce heralded a new era by the invention of integrated circuitry [4]. Around the same time, D. Kahng and M. M. Atalla successfully build the field-effect transistor that Lilienfeld proposed, based on silicon dioxide thermally grown on top of a crystalline silicon substrate. A last important step was achieved in 1963 with the employment of the complementary metal-oxide-semiconductor (CMOS) technology by Frank Wanlass [5]. The combination of n- and p-type transistors greatly improved the static power consumption, packing density, performance and manufacturing costs. From that point on, the transistor quickly spread, replacing existing technologies.

Since then, the dynamic of semiconductor market has been well described by Moore's law, which states that the number of transistors in integrated circuits doubles approximately every 18 - 24 months [6, 7], as shown in Fig. 1.1. To achieve this, the transistor dimensions are scaled down to fit more transistors to the same area and increase the functionality per area. The trend continued for over four decades as it was feasible to scale down the poly-Si/SiO_2 technology until physical limits were approached [8, 9]. At the beginning of the 21^{st} century, undesired high gate leakage currents through the thin SiO_2 insulating barrier forced the semiconductor industry to search for new alternative gate dielectrics with higher permittivity values in order to obtain comparable gate oxide capacitances with higher oxide thicknesses. However, before this was possible necessary technical and physical boundary conditions needed to be satisfied. New materials must be compatible with the existing technology and withstand all in the process involved annealing cycles without major changes in its structure. More importantly, the material must have a low reactivity with its surroundings [10]. In addition, the gate electrode material must be adjusted to account for the different band offset of the high-k material in order to achieve adequate threshold voltages. The idea of substituting SiO_2 was to attain a higher gate control with simultaneous lower gate leakage currents. However, SiO_2 was already considered to be the perfect insulator with its large band-gap and beneficial band offset constituting high barriers for electron and hole conduction. The utilization of higher permittivity materials is connected to a lower band-gap with penalties for the gate leakage current [11].

Several potential successors, e.g. Al_2O_3, TiO_2, ZrO_2 and HfO_2 were under discussion [13]. All major concerns were finally resolved by the introduction of HfO_2 in combination with a metal gate electrode. Still, SiO_2 was not entirely replaced. Because of the excellent interface quality to the Si channel and outstanding insulating properties, a thin SiO_2 buffer layer was still mandatory. The dual layer stack of SiO_2/HfO_2 guaranteed further applicability of scaling laws and provided further verification of Moore's law. The effective oxide thickness could be reduced down to levels unfeasible for a single SiO_2 gate dielectric. The first high-

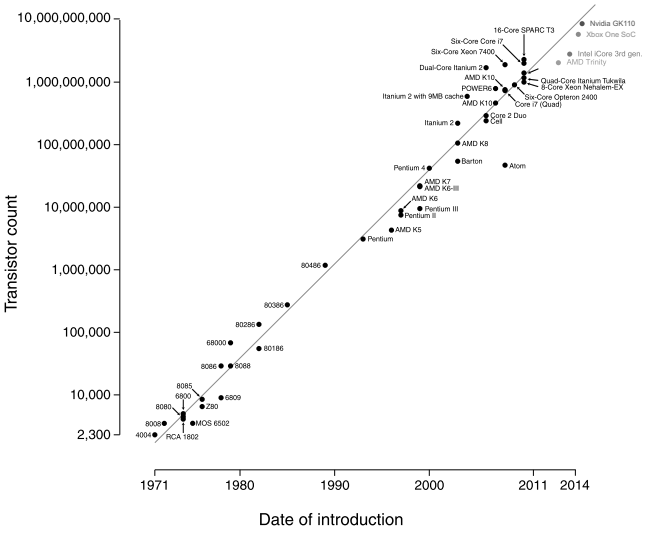


Figure 1.1.: *Moore's law shows the transistor count development through the years. His famous law predicts, that the number of transistor on a die doubles every two years which is in accordance to the microprocessor releases [12].*

k/metal gate (HKMG) technology was commercially implemented at the 45 nm node in 2007 [14]. GLOBALFOUNDRIES were able to extend their traditional poly-Si/SiO_2 technology to its 45 nm node and firstly employed a hafnium based gate process on silicon on insulator (SOI) at the 32 nm node in 2010 [15] and continued to fabricate the 28 nm node on bulk silicon.

Apart from the difficult process side, the HKMG technology represents a major challenge for MOSFET reliability. There are still unanswered questions regarding the physical degradation mechanisms which lead to the deterioration of the gate oxide properties [16, 17]. Currently, the reliability of HKMG stacks is under intensive investigation and the influence of the high-k dielectric on the overall degradation is not comprehensively understood. In comparison, there are still differences in opinions on the reliability and defect physics of standard SiO_2 and silicon oxynitride (SiON), which have been in use for the last 40 years [18]. In that sense, the altered reliability of the novel devices provides plenty of discussion for the upcoming years. As a prominent feature, the positive bias temperature instability (PBTI) evolved as a new reliability issue, which was non-existent for the poly-Si/SiO_2 era.

New phenomena such as the strong trapping and detrapping characteristics of the gate oxide also raise several questions:

- What is the origin of the PBTI degradation?
- What will be the most limiting degradation mechanism for HKMG devices?
- Are the traditional conducted characterization procedures still justified and able to capture the degradation process?

- Are constant voltage stress measurements still representative for the device degradation under realistic CMOS operating conditions?

- Can certain process treatments improve the high-k/metal gate reliability?

The upcoming sections are structured as follows: A review of the recent knowledge about MOSFET scaling effects, the different degradation mechanisms of SiO_2 based and high-k metal gate based MOSFETs and standard measurement techniques is given in chapter 2.1 - 4. The different degradation mechanisms are investigated using state of the art n- and pMOS SRAM transistors and presented in chapter 5. The most severe degradation mechanism is identified and further investigated in chapter 6 regarding the voltage and temperature acceleration of the threshold voltage and gate leakage current increase. The second to last question is addressed by taking the dynamic operation pattern of static random access memory cells as a model and to investigate the single device reliability under realistic operating conditions. Thereby, the influence on the threshold voltage shift is studied in chapter 7 and in chapter 8 regarding the dielectric breakdown. In the last chapter 9, further improvements of the gate stack properties by fluorine incorporation are discussed in terms of their benefits and impact on the overall reliability.

2. Fundamentals

The demand for better computational power with the simultaneous request for mobility and longer battery life is the driving force for MOSFET scaling. The International Technology Roadmap for Semiconductors (ITRS) describes the technological requirements for a continuous MOSFET scaling, as can be seen in table 2.1. The roadmap forecasts, that the 10 nm node will be reached by the end of the decade. However, the table also shows that manufacturable solutions are yet unknown. The ITRS predicts the gradual end of the planar technology and introduction of multi-gate CMOS transistors, which relaxes the equivalent oxide thickness (EOT) requirements. This chapter gives an overview about the general scaling rules that need to be obeyed and focuses on the different short channel effects that jeopardize the device performance and energy efficiency.

Table 2.1.: *International Technology Roadmap for Semiconductors 2012: high-performance logic device requirements [19].*

year of production			2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
MPU/ASIC Metal 1 ½ pitch		nm	38	32	27	24	21	18.9	16.8	15	13.4	11.9
physical gate length	L_g	nm	24	22	20	18	17	15.3	14	12.8	11.7	10.6
power supply voltage	V_{dd}	V	0.9	0.87	0.85	0.82	0.8	0.77	0.75	0.73	0.71	0.69
equivalent oxide thickness	EOT	nm										
planar bulk			0.88	0.84	0.79	0.73	0.67	0.61	0.55			
multi-gate				0.92	0.88	0.84	0.8	0.76	0.72	0.68	0.65	0.62
subthreshold leakage	$I_{ds,leak}$	nA/µm	100	100	100	100	100	100	100	100	100	100
saturation threshold voltage	V_t	mV										
planar bulk			285	289	296	302	306	310	312			
multi-gate				206	206	207	212	217	220	223	224	225
electrical equivalent oxide thickness	$t_{ox,inv}$	nm										
planar bulk			1.2	1.16	1.1	1.04	0.98	0.92	0.86			
multi-gate				1.32	1.28	1.24	1.2	1.16	1.12	1.08	1.05	1.02
NMOS drive current at V_{dd}	$I_{d,sat}$	µA/µm										
planar bulk			1.32	1.67	1.422	1.496	1.582	1.67	1.775			
multi-gate				1.469	1.52	1.573	1.628	1.685	1.744	1.805	1.858	1.916
NMOS dynamic power indicator	CV^2	fJ/µm										
planar bulk			0.76	0.68	0.63	0.57	0.53	0.49	0.45			
multi-gate				0.57	0.52	0.47	0.42	0.38	0.34	0.31	0.28	0.25
NMOS intrinsic delay	CV/I	ps										
planar bulk			0.64	0.57	0.52	0.47	0.42	0.38	0.34			
multi-gate				0.45	0.4	0.36	0.32	0.29	0.26	0.24	0.21	0.19

manufacturable solutions exist, and are being optimized
manufacturable solutions are known
manufacturable solutions are NOT known

2.1. Transistor scaling laws

From a general point of view, the scaling can either be done by a constant field or constant voltage scaling. For the former, the MOSFETs dimension as well as the supply voltage decrease by a certain scaling factor λ to obtain a constant electric field across the gate oxide. For the purpose of a better compatibility to older technologies and most important faster