ANGEWANDTE PHYSIK

Mikroelektronik: kein Ende der Skalierung in Sicht

Damit sich das Mooresche Gesetz weiter fortschreiben lässt, benötigt die Siliziumtechnologie neue Materialien und neue Konzepte.

Armin Tilke, Klaus Schrüfer und Chris Stapelmann

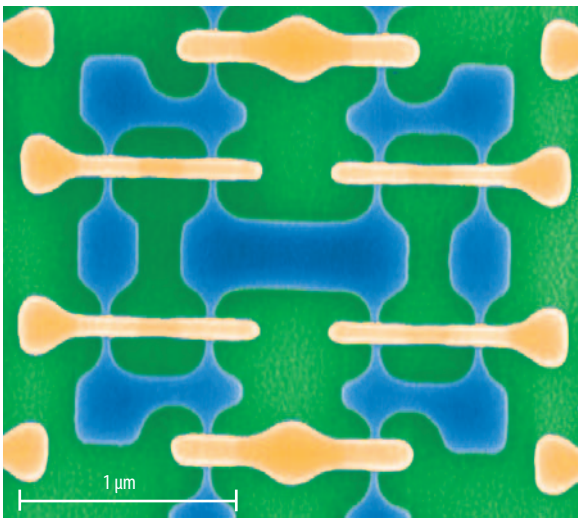
In den letzten Jahrzehnten folgte die Siliziumtechnologie ohne signifikante Verzögerungen dem Mooreschen Gesetz: Die elektronischen Schaltkreise wurden verkleinert und damit ihre Geschwindigkeit entsprechend erhöht. Inzwischen trifft die Strukturverkleinerung allerdings an physikalische Grenzen, sodass neue Materialien sowie innovative Integrations- und Bauelementkonzepte notwendig sind, um das Mooresche Gesetz weiterhin zu erfüllen.

Im Jahre 1965 machte Gordon E. Moore die Beobachtung, dass sich seit der Erfindung des planaren Transistors im Jahre 1959 bei Fairchild Semiconductor die Anzahl der Komponenten pro Chip alle zwölf Monate ungefähr verdoppelt hatte. Darauf basierend wagte Moore die Voraussage, dass sich auch in den folgenden zehn Jahren „die Komplexität der Chips oder Schaltkreise bei minimalen Komponentenkosten“ weiterhin jährlich verdoppeln würde [1].

Drei Jahre nach dieser Vorhersage verließen Moore und sein Kollege Robert Noyce Fairchild Semiconductor und gründeten Intel (Integrated Electronics Corporation). Im Jahr 1975 revidierte Moore seine ursprüngliche Feststellung dahingehend, dass sich in den folgenden Jahren die Zahl der Komponenten pro integriertem Schaltkreis in etwa 24 statt 12 Monaten verdoppeln würde [2]. Diese Aussage bezeichnete später Moores Freund und Kollege Carver Mead, Professor am Caltech, als Mooresches Gesetz (**Abb. 1**).

In den 80er-Jahren galt das Mooresche Gesetz als Schlagwort für die Verdopplung der Anzahl von Transistoren pro Chip alle 18 Monate. Im darauffolgenden Jahrzehnt stand es dagegen für die Verdopplung der Rechenleistung von Mikroprozessoren in einem festen Zeitraum und bei vergleichbaren Kosten. Tatsächlich folgte die Halbleiterindustrie in den letzten drei Jahrzehnten sehr erfolgreich der immer wieder modifizierten Mooreschen Vorhersage: Jahr für Jahr wurden Halbleiterchips mit immer höheren Schaltgeschwindigkeiten bei immer geringerem Energieverbrauch entwickelt.

Seit 1992 veröffentlichen die Halbleiterfirmen aus den USA, Europa, Japan, Korea und Taiwan die „International Roadmap for Semiconductors“ (ITRS)[1]. Dieser „Fahrplan“ enthält einen abgestimmten Blick auf die zukünftigen Trends der Halbleiterindustrie sowie der Transistorentwicklung und benennt die technischen Herausforderungen, deren Bewältigung schließlich dazu führt, dass das Mooresche Gesetz weiterhin erfüllt wird.



IMEC / Infineon

Dreidimensionale Transistoren sind aussichtsreiche Konzepte mit vielen Vorteilen gegenüber lateralen Strukturen. Dieser Multi Gate-FET ist Teil eines SRAM-Speichers, wie er als schneller Cache-Speicher für Computer verwendet wird (blau: aktives Transistorgebiet, gelb: Transistorgate, grün: Isolationsoxid).

In der Siliziumtechnologie wurde dieser Fortschritt bislang maßgeblich durch die geometrische Verkleinerung (Skalierung) der Metall-Oxid-Feldeffekttransistoren (MOSFET, siehe Infokasten) erreicht. Das Grundprinzip eines Silizium-MOSFET

1) siehe www.itrs.net

KOMPAKT

- Die Verkleinerung der Transistoren stößt bei Dimensionen unterhalb von 50 nm an physikalische Grenzen, die sich in sog. Kurzkanaleffekten oder einem zunehmenden Leckstrom durch das Gate aufgrund von Ladungsträgertunneln zeigen.
- Um die Transistorgeschwindigkeit weiter zu erhöhen, werden daher neue Konzepte entwickelt. Eine Möglichkeit besteht darin, die Beweglichkeit der Ladungsträger durch verspanntes Silizium zu verbessern, eine andere in der Verwendung von Materialien mit hoher Dielektrizitätskonstante, welche die Gate-Kapazität vergrößern.
- Insbesondere dreidimensionale Transistoren weisen gegenüber herkömmlichen Konzepten viele Vorteile auf.

Dr. Armin Tilke, Infineon Technologies Dresden, Königsbrücker Str. 180, 01099 Dresden
Dr. Klaus Schrüfer, Infineon Technologies, Am Campeon 1–12, 85579 Neubiberg
Dipl.-Ing. Chris Stapelmann, Infineon Technologies Belgien, Kapeldreef 75, B-3001 Leuven, Belgien

 1617-9437/07/0606-35

2) Der Substratsteuerfaktor beschreibt den Einfluss des Substratpotentials auf die Einsatzspannung.

(**Abb. 2**) in CMOS-Technologie, welches in einem Gate aus Polysilizium, SiO_2 als Gate-Oxid sowie selbstjustierter Source-Drain-Implantation besteht, wurde bislang kaum verändert. Bei der selbstjustierten Source-Drain-Implantation werden die Source-Drain-Gebiete ohne lithographische Strukturierung nach Ausbildung des Polysilizium-Gates implantiert. Dabei gelangt Dotierstoff sowohl in die Source-Drain-Gebiete als auch in das Gate des Transistors. Da der Kanalbereich des Transistors durch das Gate abgedeckt ist, wird er somit nicht implantiert.

In der komplementären MOS-Technologie (CMOS) sind sowohl n-Kanal-Transistoren (n-MOS) als auch p-Kanal-Transistoren (p-MOS) vorhanden, die für den Stromtransport durch Elektronen bzw. Löcher sorgen. Da beispielsweise in einer Inverterschaltung immer entweder der n-MOS- oder der p-MOS-Transistor im Sperrzustand ist und daher kaum statischer Verluststrom fließt, zeichnet sich diese Technologie durch eine sehr geringe Verlustleistung aus. Die verschiedenen CMOS-Generationen werden Knoten genannt; üblicherweise unterscheiden sich aufeinanderfolgende CMOS-Knoten in den lateralen Strukturabmessungen um den Faktor 0,7, da sich damit die Fläche halbiert.

Ein Maß für die Güte eines digitalen CMOS-Transistors ist der Drain-Sättigungsstrom $I_{D,sat}$, also der maximale Strom durch den Transistor bei gegebener Gate-zu-Source-Spannung V_G. Für Bauelemente wie z. B. Prozessoren, die bei hohen Frequenzen arbeiten, müssen die verwendeten CMOS-Transistoren schnell schalten. Da die Geschwindigkeit des Schaltvorgangs primär durch das Umladen von unvermeidbaren (parasitären) Kapazitäten bestimmt wird, vergrößert ein hoher Drainsättigungsstrom die Geschwindigkeit und damit die Transistorgüte. Der Drainsättigungsstrom $I_{D,sat}$ wird insbesondere beeinflusst durch die effektive Elektronen- bzw. Löcherbeweglichkeit μ_{eff}, die Gate-Oxidkapazität C_{gate}, die Bauelementabmessung bzw. Gate-Länge l_{gate} sowie den Substratsteuerfaktor[2] m.

$$I_{D,sat} \sim \mu_{eff} C_{gate} \frac{1}{l_{gate}} \frac{(V_G - V_{th})^a}{2m}$$

Dabei ist V_{th} die Einsatzspannung des Transistors. Die Güte eines CMOS-Transistors lässt sich folglich durch diese Bauelementparameter beeinflussen. Bisher wurden vor allem die Gate-Länge und die Gate-Oxid Dicke (T_{GOX}) verringert (klassische Skalierung), wodurch C_{gate} und damit die kapazitive Kopplung zum Transistorkanal erhöht wird, und damit direkt proportional die Ladungsträgerdichte im Kanal und der $I_{D,sat}$. Diese Entwicklung kam in den letzten Jahren allerdings ins Stocken, da eine weitere Verkleinerung der Transistoren bei Dimensionen unter 50 nm an physikalische Grenzen stößt. Insbesondere der zunehmende Leckstrom, der durch direktes Tunneln von Ladungsträgern durch das Gate-Oxid hervorgerufen wird, verhindert die weitere Verkleinerung der Gate-Oxiddicke. Bei kürzeren Kanallängen spielt außerdem die statistische Verteilung der Dotieratome im Kristall eine immer größere Rolle: So befinden sich für Transistoren mit einer Gate-Länge unter 50 nm nur einige zehn Dotieratome im Kanalbereich. Darüber hinaus tauchen sog. Kurzkanaleffekte auf, wenn die drainseitige Verarmungszone weit unter das Gate reicht und sich dadurch die vom Drain- und Source-Kontakt ausgehenden Verarmungszonen beeinflussen (**Abb. 2**). Diese Effekte verändern das elektrische Verhalten eines Transistors, sodass dieser bereits bei zu kleinen Gate-Spannungen einschaltet und nicht mehr korrekt arbeitet. Um sie zu begrenzen, müssen Gate-Länge und -Dicke gleichzeitig verringert werden, um ein Durchgreifen der elektrischen Felder des Drain-Kontaktes zum Source-Kontakt im ausgeschalteten Zustand des Transistor weitgehend zu unterbinden.

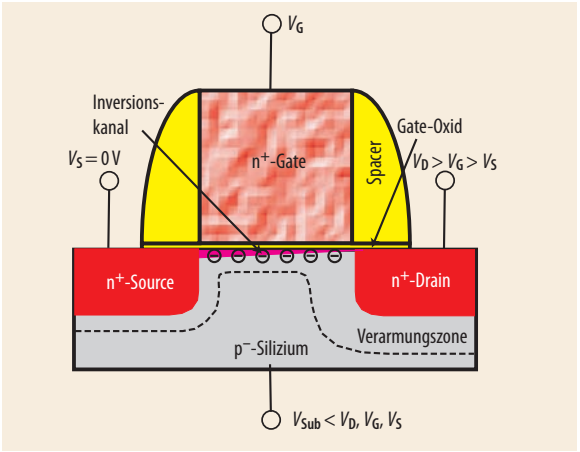


Abb. 2 Die Grundstruktur eines MOSFET (hier: n-MOS) besteht aus Source- und Drain-Kontakt sowie einem Gate, das durch ein Gate-Oxid (gelb) vom Silizium-Substrat getrennt ist. An der Grenzfläche zwischen Substrat und Gate-Oxid werden Ladungsträger induziert, die eine dünne Schicht, den sog. Inversionskanal (rot), zwischen Source- und Drain-Gebiet ausbilden. Neben der Verarmungszone unter dem Gate bilden sich auch an Source und Drain Verarmungszonen aus. Kurzkanaleffekte entstehen, wenn sich die Veramungszonen an Source und Drain zu nahe kommen. Der Spacer verhindert, dass die Source- und Drain-Dotierungen in den Kanalbereich diffundieren.

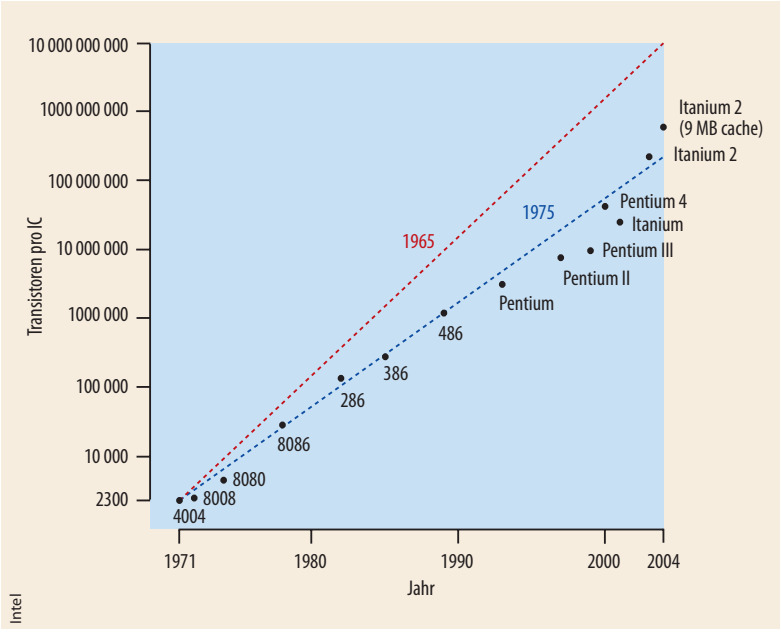


Abb. 1 Vergleich von Moores erster Voraussage aus dem Jahr 1965 und seiner Korrektur zehn Jahre später über die Verdopplung der Komplexität der Chips bzw. Schaltkreise bei minimalen Komponentenkosten.

Da dies bei konventionellem Siliziumoxid nicht länger möglich ist, konzentriert sich die Industrie inzwischen darauf, die Gate-Kapazität durch Isolatoren mit hoher Dielektrizitätskonstante k (sog. high-k-Oxide) zu erhöhen. Andere Forschungsschwerpunkte sind die Erhöhung der effektiven Ladungsträgerbeweglichkeit durch das Verspannen von Silizium sowie die Entwicklung neuer dreidimensionaler Transistorkonzepte.

Dadurch koppelt sich seit dem 90-nm-Knoten die Geschwindigkeitserhöhung der Transistoren weitgehend von der nur lateralen Strukturverkleinerung ab. Im Folgenden werden diejenigen neuen Konzepte der Transistorskalierung diskutiert, welche die Wirkung der genannten physikalischen Grenzen zu noch kleineren Strukturgrößen hinausschieben können.

Vermeiden von Kurzkanaleffekten

Die Ausdehnung der Verarmungszonen, welche in erster Näherung umgekehrt proportional zur Wurzel des Produkts aus n- und p-Dotierung am betrachteten pn-Übergang ist, lässt sich durch flache Dotierprofile sowie höhere Dotierungen begrenzen. In älteren CMOS-Technologien bewirkte eine flachere Ionenimplantation des Dotierstoffs die erforderlichen flacheren Profile. Bis zum 0,5 µm-CMOS-Knoten wurden die Strukturen zur elektrischen Aktivierung der Source- und Drain-Gebiete, also zum Einbau der Dotieratome auf reguläre Kristallgitterstellen, im Ofen erhitzt („Ofenausheilschritt"). Die lange Dauer dieses Prozesses führte allerdings zu einer starken Diffusion der Dotieratome und somit zu tiefen und stark verbreiterten Profilen. Bei neueren Technologien erhitzen intensive Quecksilberdampflampen die Siliziumscheiben für wenige Sekunden auf über 1000 °C („Rapid Thermal Annealing", RTA). Für CMOS-Knoten mit einer Gate-Länge unter 65 nm ist jedoch auch diese sehr kurze Ausheilzeit zu lang, um steile und flache Dotierprofile zu gewährleisten. Daher werden hier verschiedene, nur wenige Millisekunden dauernde Ausheilverfahren wie beispielsweise Laser- oder Blitzlampen-Technologien eingesetzt.

Beim Laserausheilen tastet ein hochenergetischer Laserstrahl die Siliziumscheibe rasterförmig ab und erhitzt dabei die Scheibenoberfläche lokal auf bis zu 1400 °C.[3)] Die Zeitkonstante der Diffusion von Dotieratomen in Silizium ist wesentlich größer als die der elektrischen Aktivierung der Dotieratome. Dank der kurzen Ausheilzeiten und der hohen Temperaturen lässt sich bei diesen neuartigen Verfahren eine höhere elektrische Aktivierung erreichen. Diese resultiert in einer größeren Leitfähigkeit der Source- und Drain-Gebiete bei gleichzeitig drastisch verringerter Ausdiffusion der Dotierstoffe (**Abb. 3**) [3].

Um das Dotierprofil flach zu halten, werden vermehrt neue Dotiertechnologien wie die Gasphasendotierung oder die Ko-Implantation verwendet. Die Ko-Implantation unterdrückt die Diffusion über

3) Der Schmelzpunkt von Silizium liegt bei 1414 °C.

DER MOS-FELDEFFEKTTRANSISTOR

Die wichtigste Struktur in der Halbleiterphysik ist der pn-Übergang in Silizium. Stößt ein mit Donatoren dotierter n-Siliziumbereich an einen mit Akzeptoren dotierten p-Bereich, diffundieren Löcher aus der p- und Elektronen aus der n-Zone in den jeweils anderen Bereich (**Abb. i**). Näherungsweise werden dabei alle freien Ladungsträger in dieser sog. Verarmungszone ausgeräumt. Die festen Ladungen der ionisierten Akzeptoren und Donatoren bilden dann ein elektrisches Feld und damit eine Spannung V_D an diesem pn-Übergang aus. Dies führt zu einem dynamischen Gleichgewicht der Ladungsträgerdiffusion. Nach der Poisson-Gleichung lässt sich daraus die Verbiegung der Bandkanten am pn-Übergang berechnen. Eine angelegte externe Spannung V verbiegt den Bandkantenverlauf weiter.

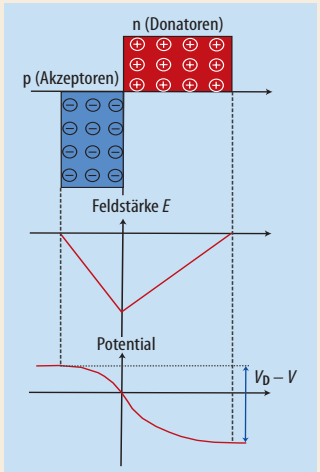


Abb. i *Stoßen ein n- und ein p-dotierter Bereich aneinander (oben), entsteht durch die entgegen gesetzten Ladungen ein elektrisches Feld (Mitte) und damit eine Spannung an diesem Übergang (unten).*

Ähnlich verhält es sich, wenn sich z. B. zwischen einem n-Gate und p-Silizium ein isolierendes Gate-Oxid befindet (Metall-Oxid-Silizium Diode). Eine ausreichend hohe, positive Spannung am Gate erzeugt eine dreiecksförmige Bandverbiegung an der Grenzfläche zwischen Gate-Oxid und p-Siliziumsubstrat. Wird das Fermi-Niveau des p-Siliziums, das nahe am Valenzband liegt, so verschoben, dass seine Energie höher ist als die des Leitungsbands (Inversion), entsteht ein dreiecksförmiger, zweidimensionaler Potentialtopf – in diesem Fall für Elektronen (**Abb. ii**). Diese können sich dann entlang der Grenzfläche frei bewegen.

Beim Feldeffekttransistor erzeugt eine Steuerspannung an der Gate-Elektrode Inversion und damit einen leitfähigen Inversionskanal im Siliziumsubstrat. Dadurch kann zwischen den Source- und Drain-Anschlüssen durch den Transistor ein Strom fließen. Steigt die Source-Drain-Spannung, dann fällt entlang des Inversionskanals durch den ohmschen Widerstand eine so hohe Spannung ab, dass am drainseitigen Ende des Kanals die resultierende Gate-Spannung keine Inversion mehr erzeugen kann. Dieser „Pinch-Off" bestimmt den Sättigungsstrom $I_{D,sat}$ des Transistors.

Die von der Drain-Elektrode ausgehende Verarmungszone überlappt teilweise mit der durch das Gate erzeugten Verarmungszone im Bereich des Inversionskanals. Dadurch entstehen Kurzkanaleffekte, die sich elektrisch in einer Verringerung von V_{th} äußern.

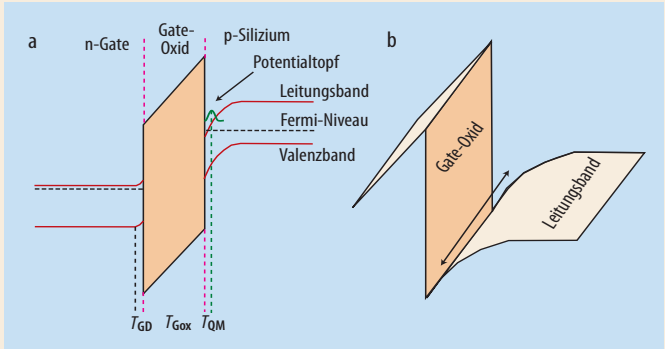


Abb. ii *Eine Spannung am Gate erzeugt eine Bandverbiegung an der Grenzfläche zwischen Gate-Oxid und p-Siliziumsubstrat (a). Die Ladungsträger können nur in der Ebene fließen, die durch den dreiecksförmigen Potentialtopf an der Oberfläche des Siliziumsubstrates gebildet wird (b).*

4) Bor wird üblicherweise als Source- und Drain-Dotierung für den p-MOS verwendet, Phosphor entsprechend für den n-MOS. Für Arsen, das auch für den n-MOS verwendet wird, hat sich die Koimplantation nicht bewährt.

Zwischengitteratome, die schon bei relativ niedrigen Temperaturen ab 300 °C einsetzt. Insbesondere die Ko-Implantation von Kohlenstoff und Fluor ist für die Unterdrückung der Diffusion von Bor und Phosphor aussichtsreich.[4)]

Allerdings lässt sich auch durch diese technologischen Neuerungen eine physikalische Grenze der Transistorskalierung nicht umgehen: Durch die höheren Dotierungen am Drain-Kontakt des Transistors werden nach dem Poisson-Gesetz die elektrischen Feldstärken an diesem pn-Übergang proportional zur Dotierung erhöht. Für den typischen Fall, dass in einem Transistor mit Gate-Länge unterhalb 50 nm innerhalb von weniger als 10 nm ein Spannungsabfall von 1 V zwischen Drain- und Kanalgebiet stattfindet, liegen die elektrischen Feldstärken bei 2–3 MV/cm. Bei diesen hohen Feldern setzt Tunneln zwischen Leitungs- und Valenzband der jeweiligen Transistorbereiche ein, das zu hohen Leckströmen am Drain-Kontakt führt. Der Übergang zu neuen Transistorkonzepten, wie sie am Ende dieses Artikels beschrieben werden, vermeidet diese Effekte.

Ladungsträgertunneln

Bei der konventionellen Skalierung erniedrigt sich Versorgungs- und damit Gate-Spannung. Dies erfordert eine höhere Gate-Kapazität, um unveränderte elektrische Feldstärken im Inversionskanal des CMOS-Transistors zu erreichen. Zudem setzt die gleichzeitige Skalierung von Gate-Länge und elektrischer Gate-Oxiddicke ebenfalls eine Verringerung der Gate-

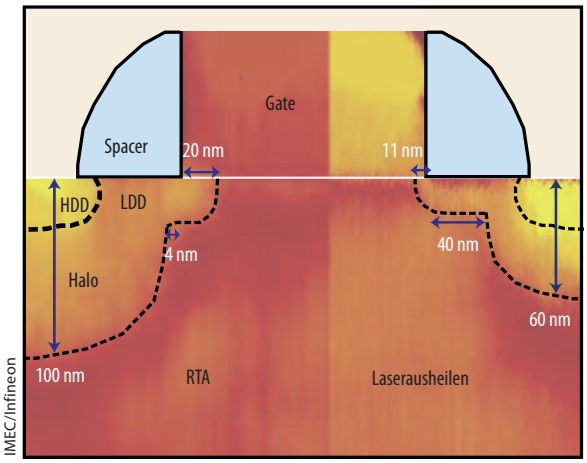


Abb. 3 Der Vergleich der Dotierprofile eines modernen CMOS-Transistors für eine konventionelle RTA-Temperung (links) sowie für das Laserausheilen (rechts) zeigt interessante Unterschiede: Die Profile direkt nach der Implantation sind praktisch identisch. Der sog. lightly doped drain (LDD)-Implant findet vor der Spacerabscheidung statt und soll nicht unter das Gate diffundieren, der highly doped drain (HDD)-Implant findet nach der Spacerausbildung statt und soll nicht weit unter den Spacer diffundieren. Mit dem Laserausheilen gelingt das deutlich besser als mit dem RTA. Ein Haloimplant unterdrückt Kurzkanaleffekte und wird unter den LDD-Implant implantiert, um dort lokal die pn-Diode höher zu dotieren. Dadurch breitet sich die Raumladungsdichte weniger weit nach unten aus .

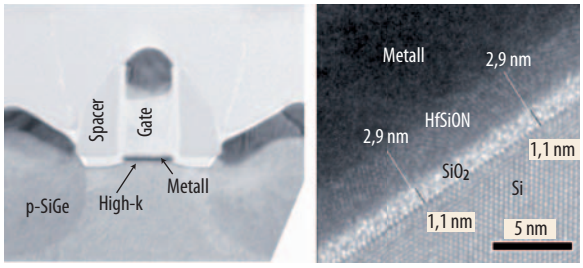


Abb. 4 Die Kombination aus high-k Gate-Isolation und einem Metall-Gate (links) erscheint für Transistoren äußerst vielversprechend. Die hochauflösende TEM-Aufnahme zeigt das high-k-Material sowie das Metall-Gate des Transistors (rechts).

Oxiddicke voraus. Für SiO_2-Filme mit einer Dicke von mehr als 2 nm ist der Leckstrommechanismus bei üblichen Gate-Spannungen von 1,5–3 V durch sog. Fowler-Nordheim-Tunneln bestimmt. Dieses setzt erst dann ein, wenn die Energiebarriere in der SiO_2-Schicht durch ein hohes angelegtes elektrisches Feld stark verbogen ist. In diesem Fall tunneln die Ladungsträger durch die dreiecksförmig verbogene Energiebarriere. Bei abnehmender Oxid-Dicke können die Ladungsträger allerdings das gesamte Gate-Oxid durchtunneln (direktes Tunneln) [4]. Dadurch wird der Verluststrom des Transistors durch das Gate-Oxid auch bei kleinen Spannungen sehr hoch. Aus diesem Grund erscheinen seit mehreren Jahren high-k-Oxide vielversprechend, die es erlauben, bei erhöhter physikalischer Oxidschichtdicke und damit stark verringertem Tunnelstrom dasselbe elektrische Feld im Inversionskanal des CMOS-Transistors zu erreichen wie mit dünnem SiO_2.

Die effektive Gesamtdicke des Gate-Oxids setzt sich aus drei Termen zusammen: Neben der physikalischen Oxidschichtdicke T_{GOX} muss bei Oxiddicken von 10–20 Å auch die Weite T_{GD} der isolierenden Verarmungszone im Polysilizium-Gate berücksichtigt werden. Ein weiterer unvermeidbarer Beitrag zur Erhöhung der effektiven Gesamtdicke besteht darin, dass sich der quantenmechanisch bestimmte Schwerpunkt der Inversionsladung nicht direkt unter der Substratoberfläche befindet, sondern um einige Angström im Abstand T_{QM} entfernt liegt. Gelangt die Gate-Oxiddicke in die Größenordnung der beiden Zusatzterme T_{QM} und T_{GD}, können diese nicht länger vernachlässigt werden (siehe Infokasten, **Abb. ii**).

In Metallen ist die Verarmungszone wegen der weitaus höheren Ladungsträgerkonzentration verglichen mit dotierten Halbleitern um mehrere Größenordnungen kleiner. Daher wird intensiv daran gearbeitet, das Gate-Material durch ein Metall oder Silizid mit metallähnlichen Eigenschaften zu ersetzen.

Aus technologischer und physikalischer Sicht sind die Hindernisse zum Einsatz von high-k-Materialien sehr hoch: Aufgrund der hohen Dichte eingebauter Ladungen bzw. Störstellen ist die Grenzfläche zwischen high-k-Material und Silizium verglichen mit SiO_2/Si elektrisch deutlich schlechter. Bei Verwendung eines Polysilizium-Gates wird das chemische Potential daher an dieser Grenzfläche fixiert („gepinnt“). Um dies und die Verarmung des Gates zu vermeiden, werden high-k-Ma-

terialien inzwischen nur noch in Kombination mit Metall- oder komplett silizierten Gates diskutiert (**Abb. 4**). Dabei sind zwei verschiedene Metalle für den n- bzw. p-MOS erforderlich, um die für beide Transistortypen unterschiedlichen Austrittsarbeiten zu erreichen.

Hafniumdioxid mit k-Werten zwischen 20 und 25 sowie plasmanitridierte Hafniumsilikate HfSiO(N) mit $k \approx 15$–20 je nach Hf- bzw. N-Gehalt sind die vielversprechendsten high-k-Materialien. Im Vergleich zu SiO_2 verringert sich der Gate-Leckstrom bei diesen Materialien etwa um den Faktor 100–1000.

Trotz der dargestellten Schwierigkeiten beim Einsatz von high-k-Dielektrika zeichnen sich mittlerweile technologische Durchbrüche ab. So haben z. B. die Halbleiterfirmen Intel und IBM angekündigt, für einige ihrer 45 nm-CMOS-Technologien schon dieses Jahr high-k-Materialien einzusetzen.

Beweglichkeitserhöhung durch Verspannung

Schon seit mehreren Jahren wird versucht, durch eine größere effektive Ladungsträgerbeweglichkeit μ_{eff} die Transistorgüte unabhängig von der Transistorskalierung zu erhöhen. Am einfachsten lässt sich dies durch das Anlegen einer Zug- oder Druckspannung im Bereich des Inversionskanals erreichen. Im Drude-Modell ist die Ladungsträgerbeweglichkeit gegeben durch

$$\mu_{\text{eff}} = \frac{q\tau}{m_{\text{eff}}} \propto I_{\text{D,sat}}$$

mit der Streurate $1/\tau$ für Elektronen bzw. Löcher, deren Ladung q sowie deren effektiver Masse m_{eff}. Die effektive Masse ist umgekehrt proportional zur Krümmung der Energiedispersionsrelation der Ladungsträger:

$$\frac{\hbar^2}{2m_{\text{eff}}} = \frac{\mathrm{d}^2E(\boldsymbol{k})}{\mathrm{d}\boldsymbol{k}^2}$$

Die angelegte Verspannung verzerrt das Gitter im Bereich des Inversionskanals und erhöht die Ladungsträgerbeweglichkeit durch zwei Effekte: Zum einen ändert sich die Bandstruktur $E(\boldsymbol{k})$ und erniedrigt dadurch die effektive Masse m_{eff}, zum anderen hebt die durch die Verspannung eingeführte Symmetriebrechung der Kristallstruktur die sechsfache Tälerentartung im Leitungsband der Elektronen auf. Dadurch verringert sich die Streuung zwischen diesen Tälern, wobei auch die Streurate $1/\tau$ abnimmt [5].

Die Ladungsträgerbeweglichkeit erhöht sich beim p-MOS hauptsächlich durch die reduzierte effektive Masse, beim n-MOS durch die reduzierte Elektronenstreuung. Eine physikalisch vollständige Betrachtung des Einflusses von Verspannungen auf die Ladungsträgerbeweglichkeit im Halbleiter ist sehr komplex.[5)]

Verbiegt sich die Bandstruktur, führt dies neben den erwähnten Effekten zu einer Umverteilung der Ladungsträger auf die Valenz- und Leitungsbänder. Zudem spalten bei hohen Verspannungen die quantenmechanischen Leitungsbandzustände auf, wodurch ihre Energieniveaus verschoben werden. Insgesamt ergibt sich eine Veränderung der zweidimensionalen Zustandsdichte.

Zurzeit existieren im Wesentlichen zwei verschiedene Methoden, um Verspannungen an den Inversionskanal eines CMOS-Transistors anzulegen. Das erste beruht darauf, eine stark verspannte dielektrische Schicht (z. B. aus Siliziumnitrid) über die Transistoren abzuscheiden. Heutige Verspannungsnitride weisen eine Zugspannung von bis zu 2 GPa auf sowie eine Druckspannung von maximal 3,5 GPa. Bisher noch teilweise unverstanden ist die sog. „Stress-Memory"-Technik, mit der sich die verwendete Zugspannung beim n-MOS im Silizium-Kristallgitter/Polysilizium-Gate einfrieren lässt. Dabei werden die Source- und Drain-Gebiete des n-FET sowie das Polysilizium des Gates durch eine Hochdosisimplantation amorphisiert. Anschließend wird eine Schicht Siliziumnitrid mit hoher intrinsischer Verspannung aufgebracht und das amorphe, verspannte Silizium durch einen kurzen, heißen Ausheilschritt rekristallisiert. Dabei entsteht eine Verzerrung im Kristallgitter, die auch nach Entfernen der verspannten Extraschicht erhalten bleibt [6].

Das zweite Verfahren, um Verspannungen an den Kanal anzulegen, besteht darin, die Source- und Drain-Bereiche des CMOS-Transistors anzuätzen und anschließend mittels lokaler Epitaxie mit einem Material anderer Gitterkonstante (Gitterfehlanpassung) aufzufüllen (**Abb. 5**). Für den p-MOS-Transistor wird seit dem 90 nm-Knoten eine p-dotierte SiGe-Schicht epitaktisch aufgewachsen, um Verspannungen zu erzeugen. Hierbei lassen sich Germaniumkonzentrationen von über 30 % erreichen, bevor bei den notwendigen SiGe-Schichtdicken die Verspannungen abgebaut werden,

5) Mittels $\boldsymbol{k}\cdot\boldsymbol{p}$-Rechnungen lassen sich die Verkrümmungen der Bänder im reziproken Raum berechnen.

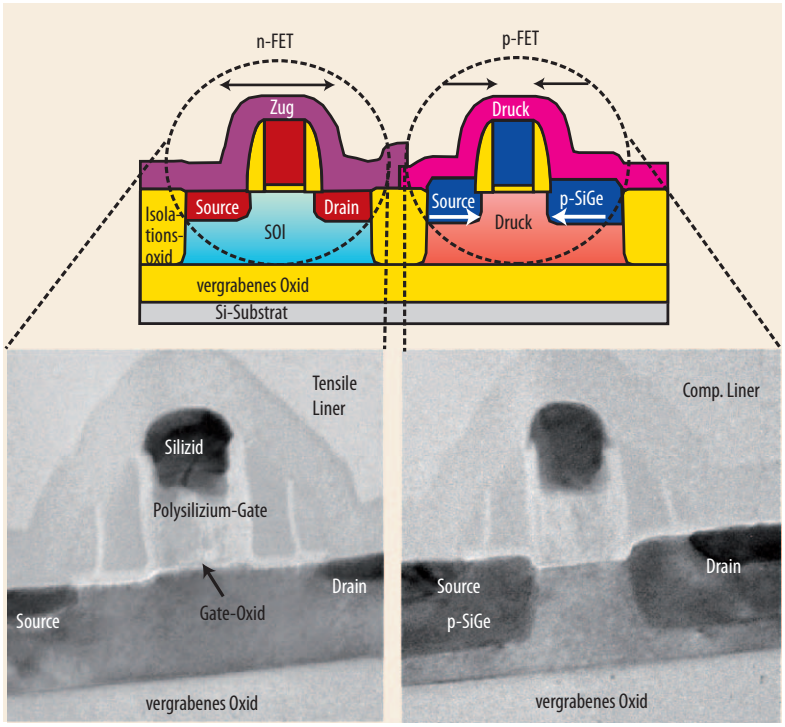
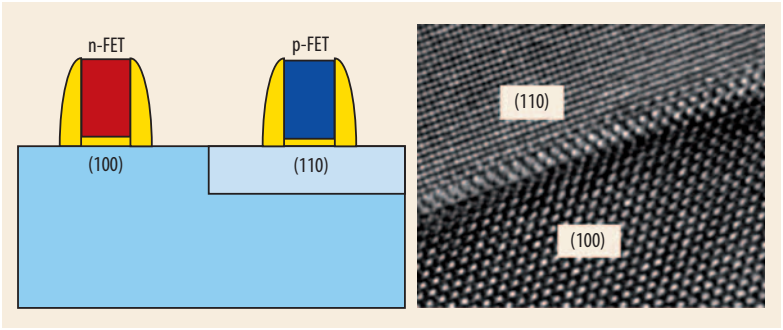


Abb. 5 An CMOS-Transistoren auf SOI-Substrat lässt sich entweder Zug- oder Druckspannung anlegen (oben). TEM-Aufnahmen zeigen die jeweiligen Strukturen (unten). Die Verspannungen werden sowohl durch Abscheiden stark verspannter Schichten (z. B. aus Siliziumnitrid, violett) erzeugt als auch durch das Anätzen der Source-Drain-Bereiche und das anschließende Auffüllen mit einem gitterfehlangepassten Material mittels lokaler Epitaxie (p-SiGe für den p-FET).



IBM

Abb. 6 Um die jeweilige Ladungsträgerbeweglichkeit zu erhöhen, ist ein Hybridsubstrat mit (110)-Kristalloberflächenorientierung für den p-FET und ein (100)-Film für den n-FET (links) erforderlich. Rechts ist das Interface zwischen den (100)- und (110)-Substraten zu sehen.

indem Versetzungen entstehen. Die für den n-MOS-Transistor verwendete SiC-Epitaxie ist dagegen technologisch noch nicht ausgereift. Zudem ist es bislang nur gelungen, Kohlenstoffkonzentrationen von bis zu 3 % einzubauen. Allerdings reicht bereits eine Kohlenstoffkonzentration von 1 % aus, um eine vergleichbar hohe Gitterverzerrung zu erzeugen wie mit der SiGe-Epitaxie bei wesentlich höherer Germaniumkonzentration von 10 %.

Im Zusammenhang mit Kristallverspannungen müssen bei zukünftigen CMOS-Knoten verstärkt die Wechselwirkungen zwischen Technologie und Schaltungsdesign berücksichtigt werden. So beeinflusst die Anzahl der durch die Verspannungsschichten (Siliziumnitrid) geätzten Kontaktlöcher zum elektrischen Anschluss der Transistoren die auf den Inversionskanal eingebrachte Gitterverzerrung. Zudem nimmt der Einfluss von Sekundärverspannungen immer weiter zu, z. B. der Störeinfluss der verspannten Oxidfüllung der Isolationsgräben zwischen benachbarten Bauelementen. Damit hängt die Transistorgüte vom Transistorlayout ab, in diesem Fall vom Abstand des Inversionskanals vom nächsten Isolationsgebiet [7].

Konventionell werden in CMOS-Technologien Silzumwafer mit einer (100)-Kristalloberflächenorientierung verwendet, da auf diesen Kristallebenen die Elektronenbeweglichkeit maximal ist. Ein neuer Ansatz der Beweglichkeitserhöhung für Löcher besteht darin, für den p-MOS-Transistor ein (110)-Silizium-Substrat zu verwenden, da in dieser Kristallebene die Löcherbeweglichkeit maximal ist. Gleichzeitig wird der n-FET in einem (100)-Substrat gefertigt (**Abb. 6**). Substrate mit zwei verschiedenen Kristallorientierungen können beispielsweise durch Waferbonden eines (110)-Filmes auf ein (100)-Substrat hergestellt werden [8]. Diese Technologie eignet sich sowohl für konventionelle Siliziumsubstrate als auch für SOI[6], allerdings ist sie so komplex, dass sie wohl zunächst nur bei Hochleistungsanwendungen wie z. B. Prozessoren eingesetzt werden wird.

Auf dem Weg zu neuen Transistorkonzepten

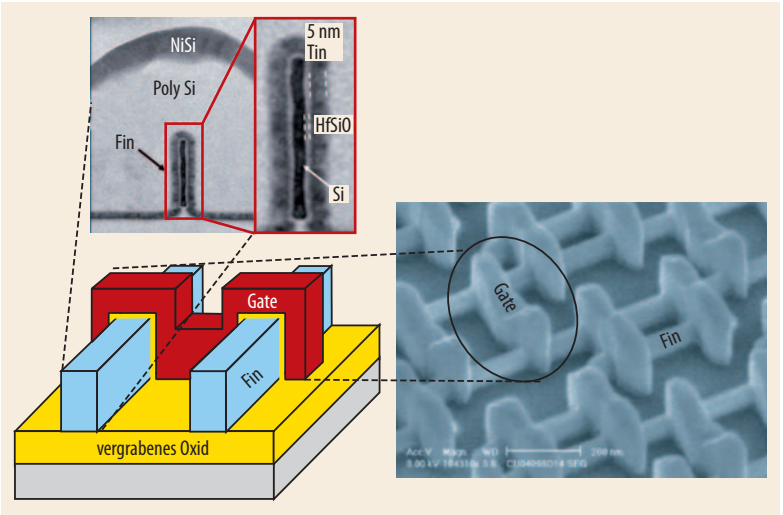
Bei kürzeren Kanallängen hat die mikroskopische Verteilung der Dotierstoffe aufgrund der geringen Anzahl von Dotieratomen einen großen Einfluss auf die Höhe der Einsatzspannung sowie deren Schwankung. Diese Schwankungen addieren sich zu den technologisch bedingten Streuungen insbesondere der Gate-Länge.

Eine weitere physikalische Grenze für die Transistorskalierung besteht darin, dass die Versorgungsspannung (Gate-Spannung) V_G verringert werden muss, um die elektrischen Felder und damit die Drain-Spannung V_D im Transistor zu begrenzen. Hingegen sinken die Einsatzspannungen V_{th} der CMOS-Transistoren nicht im gleichen Maß, da eine zu geringe Einsatzspannung den Leckstrom I_{off} im ausgeschalteten Zustand des Transistors intolerabel erhöhen würde. Diese „Stand-By"-Leistung des Transistors lässt sich ausdrücken durch:

$$P_{off} = V_G \cdot I_{off} \propto V_G \cdot \exp\left(\frac{-qV_{th}}{mkT}\right)$$

Wie aus der Gleichung abzulesen ist, fällt P_{off} mit einer Steigung von $(\ln 10)(mkT/qV_{th})$, der sog. Unterschwellstromcharakteristik. Für Transistoren mit kurzen Kanallängen beträgt diese Steigung etwa 90 mV pro Dekade. Um den Leckstrom I_{off} ausreichend klein zu halten, ist daher eine minimale Einsatzspannung V_{th} von 0,4 bis 0,5 V notwendig. Bereits seit dem 90 nm-CMOS-Knoten wird die Einsatzspannung und damit auch die Gate-Spannung kaum weiter erniedrigt. Dies hat einen beträchtlichen negativen Einfluss auf die mit V_G^2-skalierende aktive Leistung der Schaltkreise, die diese im Betrieb aufnehmen.

Alle bisher gezeigten Herausforderungen der Bauelementeskalierung legen einen Konzeptwechsel der



IMEC/Infineon

Abb. 7 Bei einer FinFET-Struktur auf einem SOI-Substrat (links) umschließen die Gate-Elektroden den Kanal auf drei Seiten und verringern dadurch den Leckstrom. Mit dieser Technik sind sehr kleine Strukturen möglich, eine Finne hat in diesem Fall eine Breite von nur 10 nm. Die REM-Aufnahme rechts zeigt einen FinFET-SRAM.

6) Silicon on Insulator, hier liegt ein 20–200 nm dünner Siliziumfilm auf einem vergrabenen Siliziumoxid. SOI wird heute vor allem für Hochleistungsprozessoren verwendet.

zugrundeliegenden CMOS-Feldeffekttransistoren nahe. Der konventionelle CMOS-Transistor ist ein Oberflächenbauelement, bei dem der elektrische Stromfluss im Bereich weniger Nanometer in der Inversionsschicht an der Grenzfläche zwischen Siliziumsubstrat und Gate-Oxid stattfindet. Durch Ausbilden einer Inversionszone im gesamten Volumen in einem dreidimensionalen Bauelement würde $I_{D,sat}$ stark ansteigen und sich damit die Transistorgüte erhöhen.

Fin-FETs aus einem oder mehreren schmalen Siliziumstegen (Multi Gate-FETs, MuG-FET) stehen schon seit längerem als dreidimensionales MOS-Bauelement auf SOI oder auf konventionellem Siliziumsubstrat zur Diskussion (**Abb. 7**). Bei ihnen ist der SOI-Film lateral strukturiert, sodass die Gate-Spannung nicht nur an der Oberfläche des Bauelements, sondern auch an dessen Seitenwänden angelegt werden kann.

Solche Bauelemente haben einige Vorzüge verglichen mit konventionellen MOSFETs. So kontrolliert die Transistorgeometrie direkt die Kurzkanaleffekte und ermöglicht dadurch die weitere Verkleinerung der Kanallänge. Auch das Problem der statistischen Dotierstoffverteilung entfällt, da undotierte Kanäle verwendet werden können. Die Unterschwellstromcharakteristik ist steiler und beträgt 63 mV pro Dekade, sodass sich V_{th} und somit auch die Gatespannung verringern lässt. Darüber hinaus erhöht die (110)-Seitenwandorientierung des Fins die Löchermobilität und damit die p-MOS-Transistorgüte.

Technologisch sind diese dreidimensionalen Strukturen allerdings noch nicht ausgereift. Zudem wäre die Hemmschwelle bei den meisten großen Halbleiterfirmen wohl sehr hoch, planare Bauelemente auch nach Lösung aller technologischen Probleme durch den Fin-FET zu ersetzen. Denn da die Stege die Transistorbreite „quantisieren", wären alle Standard-CMOS-Schaltungsbibliotheken zu ändern. So werden dreidimensionale Bauelemente vermutlich erst Einsatz finden, wenn die Verwendung der planaren Bauelemente an harte physikalische Grenzen stößt, wie es beim 22 nm- oder 16 nm-Knoten der Fall sein könnte, also in etwa fünf bis sieben Jahren. Die Industrie muss sich jedoch schon heute auf diesen revolutionären Wechsel der Transistorarchitekur vorbereiten. Infineon hat bereits voll funktionsfähige komplexe Schaltungen mit mehreren 10 000 Transistoren in MUG-FET-Architektur präsentiert [9].

Ausblick

Gordon E. Moore ist fest davon überzeugt, dass Ingenieure und Wissenschaftler die heute scheinbar unbezwingbaren Skalierungsprobleme – wie schon so oft in der Vergangenheit – durch ihren Innovations- und Erfindungsreichtum überwinden werden. Aber auch wenn die Halbleitertechnologie durch neuartige Konzepte weiterhin dem Mooreschen Gesetz folgen kann, werden in Zukunft auch andere Abhängigkeiten eine zunehmende Rollen spielen: So besagt Rocks Gesetz, dass sich die Kosten für Halbleiter-Produktionsanlagen alle vier Jahre verdoppeln. Dieses oft ignorierte ökonomische Argument dürfte die Bauelementeskalierung zukünftig stärker beeinflussen als die zugrundeliegende Physik. Dass zudem nicht immer das technologisch Machbare auch das technisch oder ökonomisch beste Ergebnis bringt verbirgt sich hinter dem Begriff „More than Moore". Hierbei geht es vor allem um die Integration verschiedener Komponenten auf einem Chip. Beispielsweise erreicht Infineon durch die Integration von Hochfrequenz- und reinen CMOS-Komponenten eine Verringerung der Herstellungskosten für Mobiltelefone um bis zu 40 %, obwohl mit dem verwendeten 130-nm-Prozess noch eine vergleichsweise konventionelle Technologie zum Einsatz kommt. Zukünftig wird dieses optimierte Systemkonzept zusammen mit leistungsfähigeren 65-nm-Technologien und darunter kombiniert werden, um dann auch die Vorteile der höheren Integrationsdichte der Transistoren bei immer größeren Geschwindigkeiten und weniger Leistungsaufnahme zu nutzen.

Literatur

[1] *G. E. Moore*, Electronics **38**(8), 114 (1965)
[2] *G. E. Moore*, IEDM Tech. Dig., 11 (1975)
[3] *Z. Luo* et al., IEEE Electron Dev. Meeting, 489 (2005)
[4] *D. A. Muller* et al., Nature **399**, 758 (1999)
[5] *S. E. Thompson* et al., IEDM Techn. Dig., 221 (2004).
[6] *M. Horstmann* et al., IEEE Electron Dev. Meeting, 233 (2005)
[7] *A.T. Tilke* et al., IEEE Trans Semicon. Manufacturing, 59 (2007)
[8] *M. Yang* et al., Tech. Digest VLSI Technol., 160 (2004)
[9] *K. Schruefer* et al., VLSI-Design, Automation and Test (2007) und *K. von Arnim*, Proceedings VLSI Technology (2007)

DIE AUTOREN

Armin Tilke hat Physik an der TU München studiert und im Bereich der Silizium-Nanoelektronik am Center for Nanoscience (LMU München) promoviert. Zwischen 1999 und 2003 arbeitete er bei Infineon Technologies Dresden und München, anschließend war er für die Allianz aus IBM, Chartered, Infineon und Samsung in East Fishkill (USA) tätig. Inzwischen entwickelt er bei Infineon Technologies Dresden hochintegrierte Smart Power Bauelemente und ist dort für Innovationsthemen verantwortlich.

Klaus Schrüfer hat an der Friedrich-Alexander Universität Erlangen Physik studiert und dort 1995 im Bereich Theoretische Halbleiterphysik promoviert. Im Anschluss war er bei der Siemens AG München und Dresden u. a. für die CMOS-Transistorentwicklung zuständig. Zwischen 1998 und 2001 war er innerhalb der Allianz aus IBM und Siemens (später Infineon) in East Fishkill (NY, USA) für die Transistorentwicklung in der 0,18-µm- und 0,13-µm-CMOS-Technologie verantwortlich. Seit 2001 ist er Principal bei Infineon Technologies München für den Bereich CMOS Logik Devices und damit u. a. verantwortlich für Innovationen im Bereich neuer Materialen und Transistor-Architekturen.

Chris Stapelmann erhielt 1998 seinen Abschluss als Diplom-Ingenieur von der RTWH Aachen. Danach arbeitete er bei Novellus Systems in Portland, Oregon. Seit 2001 arbeitet er bei Infineon, zunächst als Anlagenexperte in Richmond, Virgina, dann als Prozessentwicklungsingenieur in East Fiskill, New York. Inzwischen leitet er die „Front-End-of-Line" Technologieentwicklung für Infineon am Forschungszentrum IMEC und ist On-Site Manager der Infineon-Niederlassung Belgien. Seit 2006 ist er Mitglied der ITRS.

