Halbleitertechnologie von A bis Z

Lithografie

Belichten und Belacken

Übersicht

In der Halbleiterfertigung werden Strukturen auf Siliciumscheiben mittels lithografischer Verfahren hergestellt. Dabei wird zuerst ein strahlungsempfindlicher Film, meist eine Fotolackschicht, auf dem Wafer aufgebracht, strukturiert und mit Ätzverfahren in die darunter liegende Schicht übertragen.

Die Fototechnik beinhaltet dabei folgende Prozessschritte:

- Aufbringen eines Haftvermittlers und Entfernen von Wasser auf dem Wafer
- Belacken der Wafer
- Stabilisieren der Lackschicht
- Belichten
- Entwickeln des Lackes
- Aushärten des Lackes
- Kontrolle

Bei einigen Prozessen wie der Ionenimplantation, dient der Fotolack als Schutzschicht um bestimmte Bereiche auf dem Wafer von der Implantation auszuschließen. Eine Übertragung der Lackmaske durch einen Ätzprozess findet hier nicht statt.

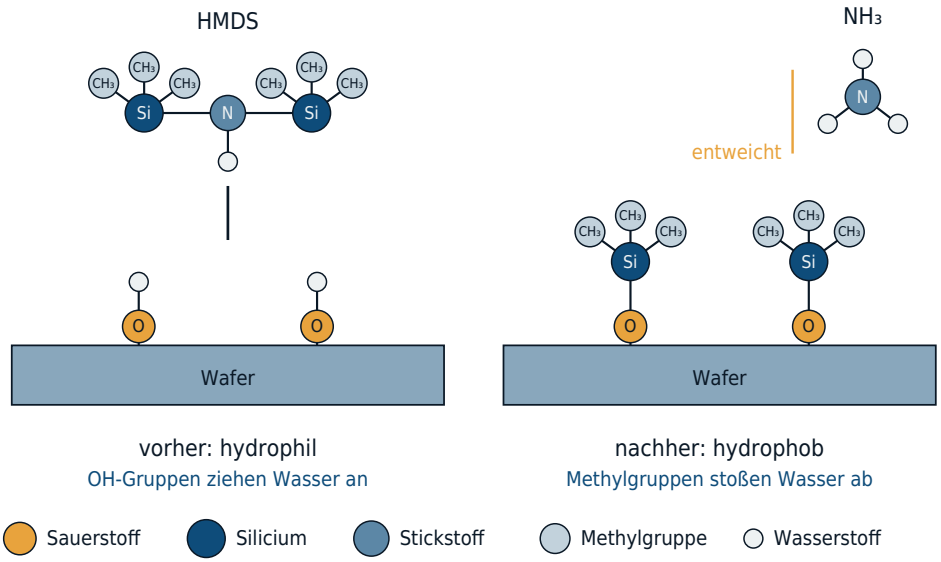
Aufbringen eines Haftvermittlers

Zu Beginn werden die Wafer gereinigt und ausgeheizt (Pre-Bake) um anhaftende Partikel zu beseitigen und angelagertes Wasser zu entfernen. Die Oberfläche der Wafer ist Wasser anziehend (hydrophil) und muss vor dem Aufbringen der Lackschicht hydrophob, also Wasser abstoßend und somit Lack anziehend gemacht werden. Dazu wird auf den Wafern ein Haftvermittler, meist Hexamethyldisilazan (HMDS), aufgebracht. Die Wafer werden dabei dem Dampf dieser Flüssigkeit ausgesetzt, so dass sich die Scheibenoberflächen damit benetzten.

Durch die Feuchtigkeit in der Umgebungsluft befinden sich auch nach dem Ausheizen immer Wasserstoff H oder Hydroxidionen OH^- an der Waferoberfläche. Das HMDS spaltet sich in Trimethylsiliciumgruppen $Si(CH_3)_3$ auf und entfernt den Wasserstoff unter Bildung von Ammoniak NH_3.

Anlagerung von HMDS

Aus der wasseranziehenden Oberfläche wird eine wasserabweisende



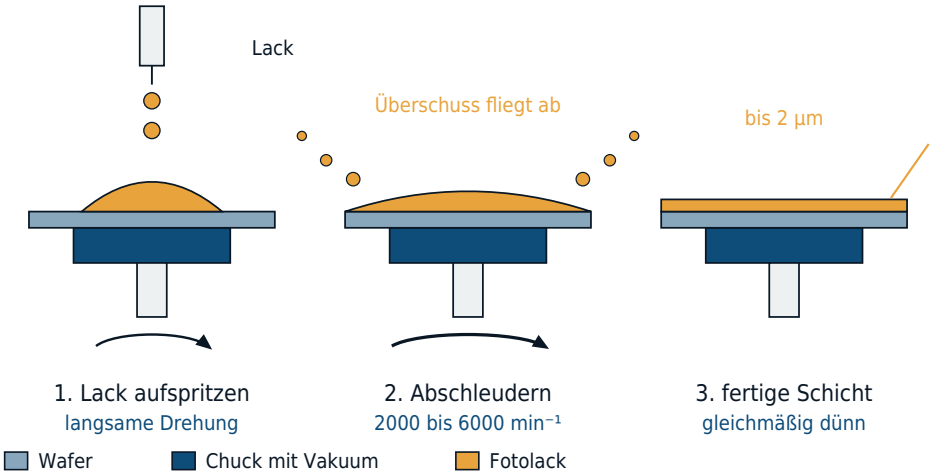
Das HMDS spaltet sich in zwei Trimethylsiliciumgruppen. Jede bindet an ein Sauerstoffatom der Oberfläche und nimmt dessen Wasserstoff mit; zusammen mit dem Stickstoff entsteht daraus Ammoniak.

Belacken

Die Belackung der Wafer erfolgt durch eine Schleuderbeschichtung auf einem drehbaren Teller mit Vakuumansaugung (Chuck). Bei niedriger Drehzahl wird Lack in der Mitte der Scheibe aufgespritzt und dann bei 2000–6000 Umdrehungen pro Minute durch die Zentrifugalkraft zu einer homogenen Lackschicht auseinander gezogen. Deren Dicke beträgt je nach anschließendem Prozess bis zu 2 µm. Die Dicke hängt dabei von der Drehzahl und der Zähigkeit (Viskosität) des Lackes ab.

Belacken durch Aufschleudern

Die Zentrifugalkraft zieht den Lacktropfen zu einer dünnen Schicht auseinander



Die Schichtdicke folgt aus Drehzahl und Zähigkeit des Lacks: Je schneller der Teller dreht und je dünnflüssiger der Lack, desto dünner die Schicht. Der weitaus größte Teil des aufgespritzten Lacks wird dabei abgeschleudert.

Damit der Lack auf der Scheibe gleichmäßig verfließen kann enthält er Wasser und Lösungsmittel, die ihn weich machen. Zur Stabilisierung der Lackschicht wird der Wafer danach bei ca. 100 °C ausgeheizt (Post-Bake bzw. Soft-Bake). Wasser und Lösungsmittel werden teilweise verdampft, eine Restfeuchtigkeit muss für die Belichtung erhalten bleiben.

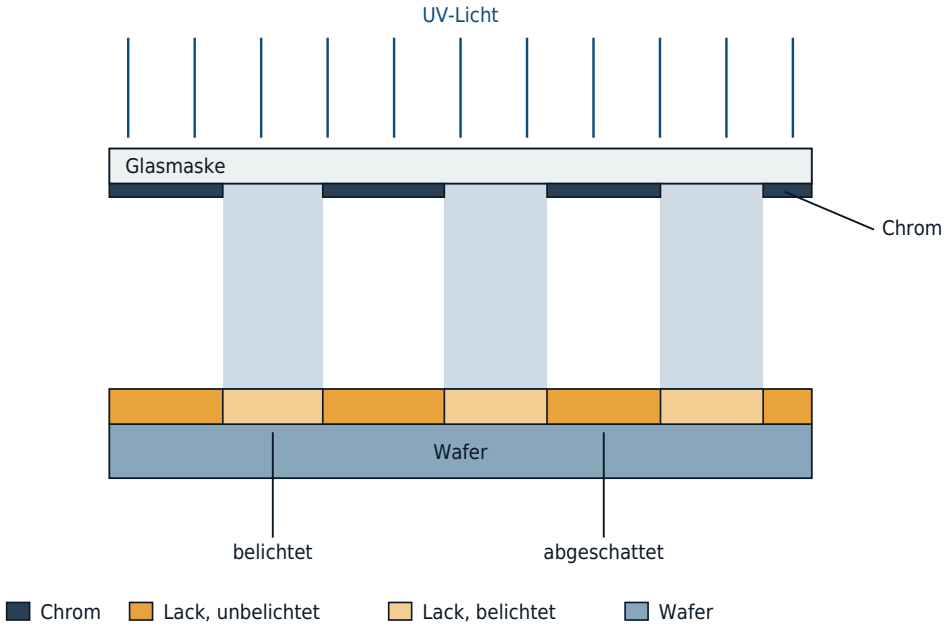
Heutzutage kommen auf Grund der geringen Strukturgrößen oft noch weitere Hilfsschichten zum Einsatz, welche entweder vor oder nach dem Fotolack auf dem Wafer aufgebracht werden. Diese Schichten können u. a. als Antireflexschicht oder zur zusätzlichen Planarisierung dienen.

Belichtung

In einer Belichtungsanlage befindet sich eine Glasmaske die teilweise mit Chrom beschichtet ist, dadurch werden partiell Bereiche des belackten Wafers belichtet und andere abgeschattet.

Belichtung durch eine Maske

Wo Chrom auf der Maske liegt, bleibt der Lack unbelichtet



Die Maske trägt das Muster als Chromschicht auf Glas. Beim Entwickeln wird je nach Lacktyp entweder der belichtete oder der unbelichtete Teil abgelöst – zurück bleibt die strukturierte Lackschicht.

Je nach Art des Lackes werden belichtete Teile löslich oder unlöslich. Mit Hilfe einer Entwicklerlösung werden die löslichen Bereiche entfernt, so dass eine strukturierte Lackschicht erhalten bleibt. Bei klassischem Positivlack der i-Linie spaltet sich ein Stickstoffmolekül N_2 durch das energiereiche UV-Licht ab. Zurück bleibt ein Keto-Karben, das sich aus energetischen Gründen zu Keten ($CH_2{=}C{=}O$) umwandelt. Unter Aufnahme von Feuchtigkeit aus der Umgebungsluft bildet sich aus dem Keten eine Carbonsäure. Bei 248 nm und darunter wird dieser Mechanismus durch chemisch verstärkte Lacke ersetzt, die eine Photosäure freisetzen (siehe Fotolack).

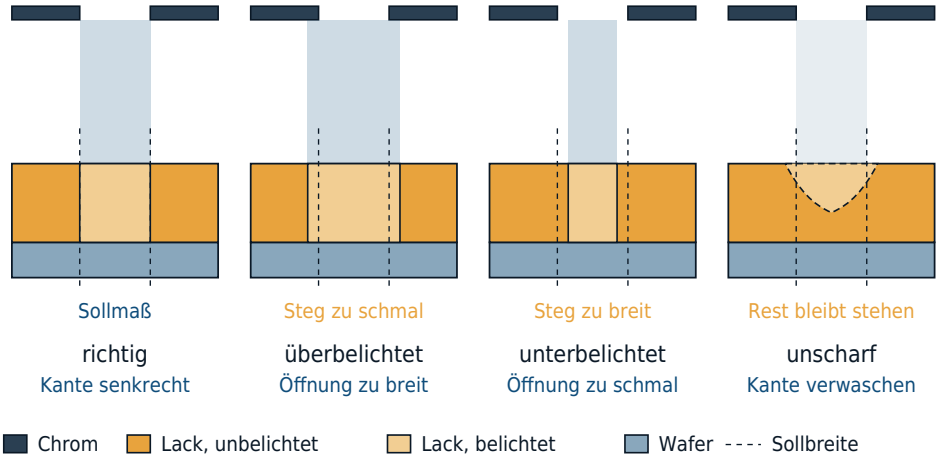
Die Belichtungszeit ist sehr wichtig damit die Strukturen die exakte Größe erhalten. Je länger die Scheibe den Strahlen der Belichtungsanlage ausgesetzt wird, desto größer werden dabei die belichteten Bereiche. Eine exakte Bestimmung der korrekten Belichtungsdauer zum Erreichen der vorgegebenen Strukturbreiten mit einem oder mehreren Testwafern (Vorläufern) ist notwendig, da sich der Lack je nach Umgebungstemperatur auch unterschiedlich verhalten kann.

Bei einer Überbelichtung werden Lackstege und damit die darunter liegenden Strukturen zu klein, Kontaktlöcher werden zu groß. Bei einer zu kurzen Belichtungszeit sind die Kontaktlöcher nicht geöffnet, Leiterbahnen sind zu breit und stehen unter Umständen miteinander in Kontakt. Zudem führt eine

schlechte Fokussierung zu unbelichteten Bereichen, so dass Kontaktlöcher bei der Entwicklung später nicht freigelegt werden und zwischen Leiterbahnen Verbindungen bestehen bleiben, die zu Kurzschlüssen führen.

Belichtungsprofile

Belichtungszeit und Fokus entscheiden über die Breite der Struktur



Je länger belichtet wird, desto breiter der belichtete Bereich: Lackstege werden schmaler, Kontaktlöcher größer. Bei zu kurzer Belichtung oder schlechtem Fokus bleibt am Boden Lack stehen und das Kontaktloch öffnet nicht.

Je nach nachfolgendem Prozess muss das Lackmaß, also die Breite der Lackstege oder die Größe der Kontaktlöcher angepasst werden. Bei isotropen Ätzungen (die Ätzung geschieht sowohl in vertikaler als auch horizontaler Richtung) wird die Maskierung nicht 1:1 in die zu strukturierende Schicht übertragen.

Eingesetzte Belichtungsverfahren

Zur Belichtung werden je nach Anforderung energiereiche Strahlen wie UV-Licht, Röntgenstrahlung, Elektronenstrahlen und Ionenstrahlen eingesetzt. Dabei gilt: Je kürzer die Wellenlänge der Strahlung, desto kleiner sind auch die erreichbaren Strukturbreiten. Den Zusammenhang beschreibt das Rayleigh-Kriterium:

$$CD=k_1\cdot\frac{\lambda}{NA}$$

Dabei ist λ die Belichtungswellenlänge, NA die numerische Apertur des Objektivs und k_1 ein Prozessfaktor, der die Beleuchtungsart, die Maskentechnik und den Fotolack zusammenfasst. Die Schärfentiefe nimmt dabei mit dem Quadrat der Apertur ab:

$$DOF=k_2\cdot\frac{\lambda}{NA^2}$$

Jede Verbesserung der Auflösung wird also mit einem kleineren Fokusfenster bezahlt – deshalb sind planare Waferoberflächen (siehe CMP) eine Voraussetzung für feine Strukturen.

Die Wellenlängen im Überblick

Wellenlänge	Quelle	Einsatz
436 nm (g-Linie) 365 nm (i-Linie)	Quecksilberdampflampe	historisch; die i-Linie wird für unkritische Ebenen, Leistungshalbleiter und MEMS weiterhin genutzt
248 nm	Kryptonfluorid-Excimerlaser	gröbere Ebenen, Analog- und Leistungstechnik
193 nm	Argonfluorid-Excimerlaser	Arbeitspferd der Fertigung, trocken und als Immersionsbelichtung
13,5 nm (EUV)	Zinn-Plasma, mit Laser gezündet	kritische Ebenen aller führenden Logik- und Speicherprozesse

Ursprünglich war als Zwischenschritt auch der Fluorlaser mit 157 nm geplant. Er erwies sich als nicht praktikabel, da die dafür nötigen Calciumfluoridlinsen Feuchtigkeit aus der Luft aufnehmen; zudem wäre eine Immersionsbelichtung mit Wasser nicht möglich, weil Wasser Licht dieser Wellenlänge absorbiert.

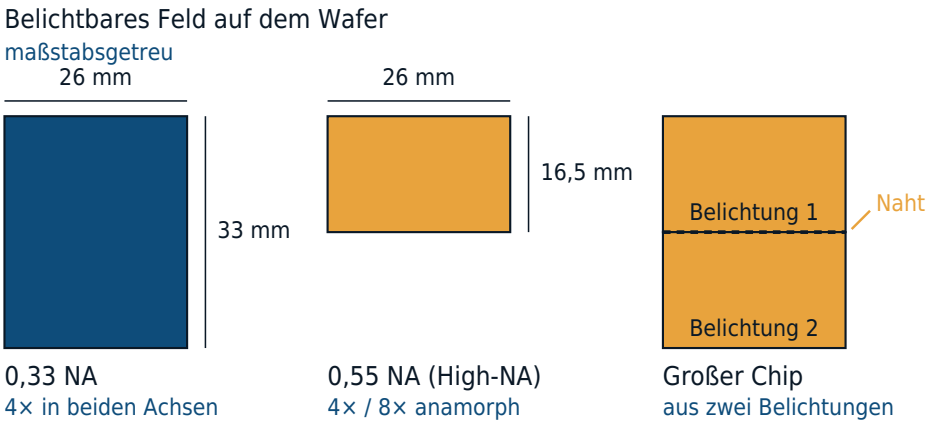
Immersionsbelichtung

Statt die Wellenlänge weiter zu verkürzen, lässt sich die numerische Apertur erhöhen. Dazu wird der Spalt zwischen letzter Linse und Wafer mit hochreinem Wasser gefüllt, dessen Brechungsindex bei 193 nm etwa 1,44 beträgt. Damit sind Aperturen bis 1,35 möglich, gegenüber rund 0,93 in Luft. Eine einzelne Immersionsbelichtung löst dabei Strukturen bis etwa 38 nm Halbperiode auf – das ist die physikalische Grenze dieser Technik. Feinere Strukturen entstehen mit 193 nm nur noch, indem man ein Muster auf mehrere Belichtungen aufteilt (siehe Multipatterning).

EUV-Lithografie

Der Sprung auf extremes Ultraviolett mit 13,5 nm ist seit 2019 in der Serienfertigung. Die Strahlung entsteht, indem Zinntröpfchen von einem Hochleistungslaser zu einem Plasma verdampft werden; da EUV von Luft und von jedem Glas absorbiert wird, arbeitet die gesamte Anlage im Vakuum und bildet ausschließlich über Spiegel ab. Die heutige Anlagengeneration mit einer Apertur von 0,33 erreicht in einer Belichtung etwa 13 nm Halbperiode und ersetzt damit an vielen Ebenen zwei oder mehr Immersionsschritte.

Belichtbares Feld bei 0,33 NA und 0,55 NA



Seit 2024 kommt die High-NA-Generation mit einer Apertur von 0,55 hinzu, die etwa 8 nm Halbperiode auflöst. Der Preis dafür ist eine anamorphe Optik: Das Bild wird in einer Richtung um den Faktor 4, in der anderen um den Faktor 8 verkleinert, wodurch sich das belichtbare Feld auf 26 × 16,5 mm halbiert. Große Chips müssen deshalb aus zwei aneinandergesetzten Belichtungen entstehen. Im Juli 2026 wurden erstmals High-NA-belichtete Ebenen in einem Serienprodukt qualifiziert; der breite Einsatz wird für 2027 und 2028 erwartet.

Weitere Verfahren

Röntgen- und Ionenstrahlen haben die Fertigung nie erreicht und sind der Forschung vorbehalten. Elektronenstrahlschreiber sind dagegen unverzichtbar – nicht zur Belichtung der Wafer, sondern zur Herstellung der Fotomasken.

Multipatterning

Eine einzelne Immersionsbelichtung mit 193 nm löst Strukturen bis etwa 38 nm Halbperiode auf. Als die Bauelemente diese Grenze unterschritten, bevor die EUV-Lithografie verfügbar war, blieb nur ein Weg: Das Muster wird auf mehrere Belichtungs- und Ätzschritte verteilt, die jeder einzeln innerhalb der Auflösungsgrenze liegen. Zusammen ergeben sie ein feineres Muster, als eine Belichtung erzeugen könnte.

Mehrfachbelichtung mit zwei Masken (LELE)

Beim Litho-Etch-Litho-Etch-Verfahren wird das Zielmuster in zwei Teilmuster zerlegt, die jeweils den doppelten Strukturabstand haben. Das erste wird belichtet und geätzt, danach folgt ein zweiter kompletter Durchlauf, dessen Strukturen genau in die Lücken des ersten fallen.

Der Nachteil liegt in der Justierung: Die Lage der zweiten Ebene zur ersten geht

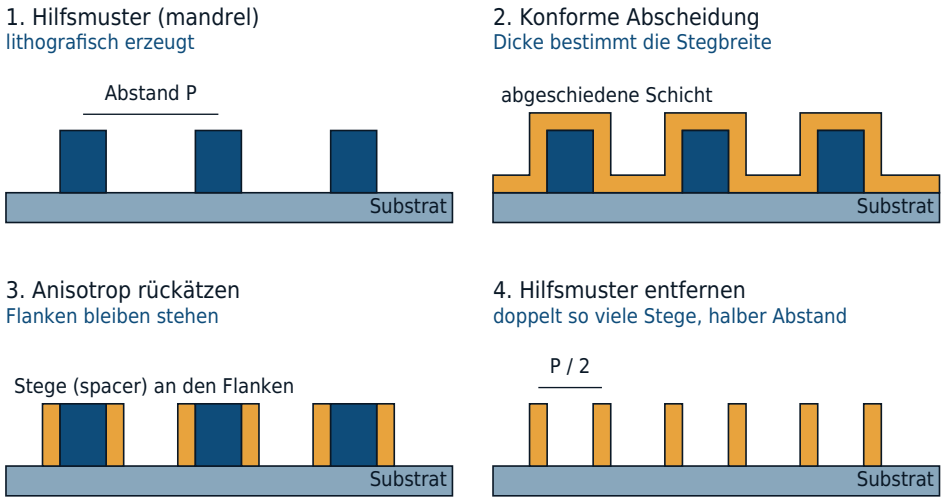
direkt in die Strukturbreite ein. Jeder Justierfehler erscheint als abwechselnd zu breiter und zu schmaler Abstand – ein systematischer Fehler, der die Bauelemente ungleichmäßig macht. Mit drei oder vier Durchläufen (LELELE und weiter) wächst dieses Problem, ebenso die Zahl der Prozessschritte und damit die Kosten.

Selbstjustierende Verdopplung (SADP)

Das selbstjustierende Verfahren umgeht das Justierproblem, indem die feinen Strukturen nicht belichtet, sondern abgeschieden werden:

- Ein Hilfsmuster (mandrel) wird mit normaler Auflösung lithografisch erzeugt.
- Darüber wird konform eine dünne Schicht abgeschieden, üblicherweise per Atomlagenabscheidung. Ihre Dicke bestimmt die späteren Strukturbreiten – und sie ist sehr viel genauer einstellbar als eine Belichtung.
- Ein anisotroper Ätzschritt entfernt die Schicht auf den waagerechten Flächen und lässt an den Flanken des Hilfsmusters Stege stehen (spacer).
- Das Hilfsmuster wird selektiv entfernt. Zurück bleiben doppelt so viele Stege, wie das Hilfsmuster Strukturen hatte.

Ablauf der selbstjustierenden Verdopplung



Da die Stege durch das Hilfsmuster selbst positioniert werden, gibt es zwischen ihnen keinen Justierfehler. Wiederholt man das Verfahren, entstehen vier Strukturen pro ursprünglicher (SAQP). Der Preis: Es entstehen immer geschlossene Ringe um das Hilfsmuster, und die Struktur ist zwangsläufig regelmäßig. Beides muss mit zusätzlichen Masken korrigiert werden, die unerwünschte Abschnitte gezielt durchtrennen (cut mask) oder Enden definieren. Das Verfahren eignet sich deshalb für dichte, regelmäßige Muster –

Leitbahnen, Gate-Reihen, Speicherzellenfelder – und nicht für beliebige Layouts.

Bedeutung heute

Multipatterning ist keine Übergangslösung geblieben. Auch mit EUV wird es weiter eingesetzt, nur an anderer Stelle: Wo EUV die kritischen Ebenen in einer Belichtung schafft, entfällt die Aufteilung; wo Strukturen unterhalb der EUV-Auflösung liegen, wird sie erneut nötig. Zugleich hat sich der Entwurf dauerhaft verändert. Layouts folgen heute strengen Regeln mit einheitlichen Rastern und einheitlichen Vorzugsrichtungen, weil nur solche Muster überhaupt zerlegbar sind – die Lithografie schreibt dem Schaltungsentwurf ihre Beschränkungen vor.

Belichtungsverfahren

Übersicht

Die Fototechnik kann, je nach Art der Bestrahlung, in verschiedene Verfahren unterteilt werden: optische Lithografie (Fotolithografie), Elektronenstrahllithografie, Röntgenstrahllithografie und Ionenstrahllithografie.

Bei der optischen Lithografie werden strukturierte Fotomasken (Reticle) verwendet. Die Belichtung mit UV-Licht oder Gaslasern erfolgt entweder im Maßstab 1:1 oder reduzierend, beispielsweise 4:1 oder 10:1.

Die optische Lithografie zerfällt heute in zwei technisch völlig verschiedene Zweige:

- DUV (deep ultraviolet, 248 und 193 nm): Die Maske ist transparent, abgebildet wird durch ein Linsensystem, die Anlage arbeitet an Luft oder mit einem Wasserfilm zwischen Objektiv und Wafer.
- EUV (extremes Ultraviolett, 13,5 nm): Da diese Strahlung von jedem Material absorbiert wird, arbeitet die Anlage im Vakuum, bildet über Spiegel ab und die Maske ist kein Transparent, sondern ein Spiegel.

Die nachfolgend beschriebenen Belichtungsverfahren sind nach steigendem Auflösungsvermögen geordnet. Kontakt- und Abstandsbelichtung finden sich heute nur noch in der Mikromechanik, in der Forschung und in der Ausbildung; die Chipfertigung nutzt ausschließlich die reduzierende Projektionsbelichtung.

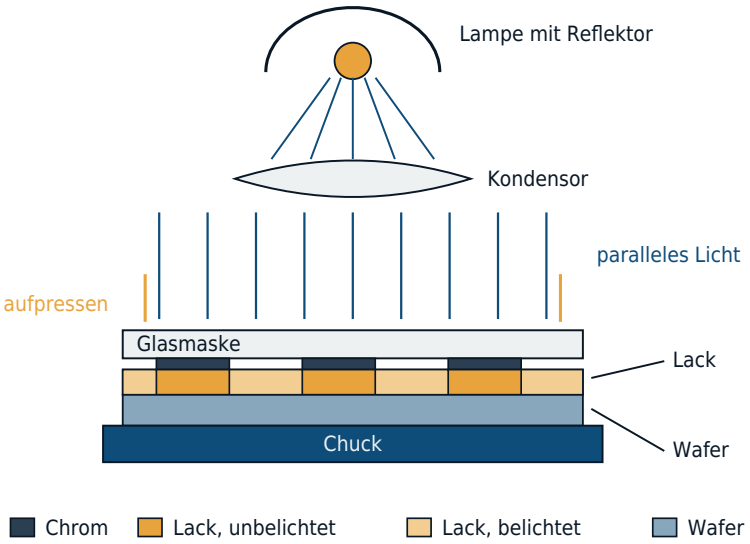
Kontaktbelichtung

Das Kontaktbelichtungsverfahren ist das älteste angewandte Verfahren. Dabei wird die Maske direkt auf die Lackschicht gepresst, die Strukturen im Maßstab

1:1 übertragen. Auflösungsbegrenzende Streu- bzw. Beugungseffekte des Lichts treten nur an den Strukturkanten auf. Die erzielten Strukturweiten sind bei diesem Verfahren jedoch gering. Da alle Chips auf einmal belichtet werden ist der Scheibendurchsatz bei dieser Technik sehr hoch, der Aufbau der Belichtungsanlagen ist relativ einfach.

Kontaktbelichtung

Die Maske liegt unmittelbar auf dem Lack, die Übertragung erfolgt 1:1



Weil Maske und Lack sich berühren, entstehen Beugungseffekte nur an den Strukturkanten. Der Aufbau ist einfach und belichtet alle Chips auf einmal. Dafür verschmutzt und verkratzt die Maske, weil sie auf den Lack gepresst wird.

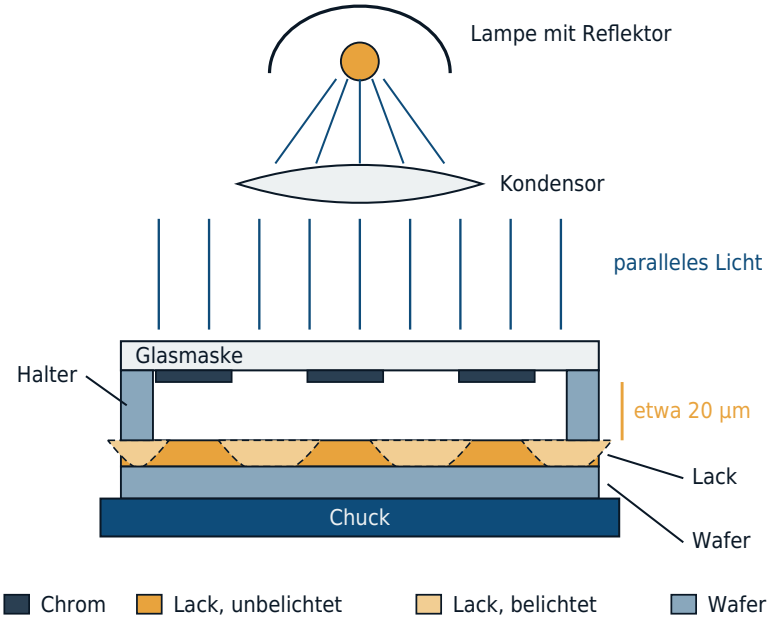
Die Nachteile liegen jedoch auf der Hand: durch die angepresste Maske auf den Lack verschmutzt diese schnell, oder kann, wie auch der Lack, dabei verkratzt werden. Befinden sich Partikel zwischen Scheibe und Maske wird die Abbildung durch den Abstand von Maske und Wafer verschlechtert.

Abstandsbelichtung

Bei der Abstands- oder Proximitybelichtung wird der Kontakt von Maske und Scheibe durch einen Abstandshalter (ca. 20 µm) vermieden. Dabei wird jedoch nur ein Schattenbild der Maske auf dem Wafer abgebildet, welches eine deutlich schlechtere Auflösung der Strukturen bietet.

Abstandsbelichtung

Abstandshalter halten die Maske über dem Lack – es entsteht ein Schattenbild



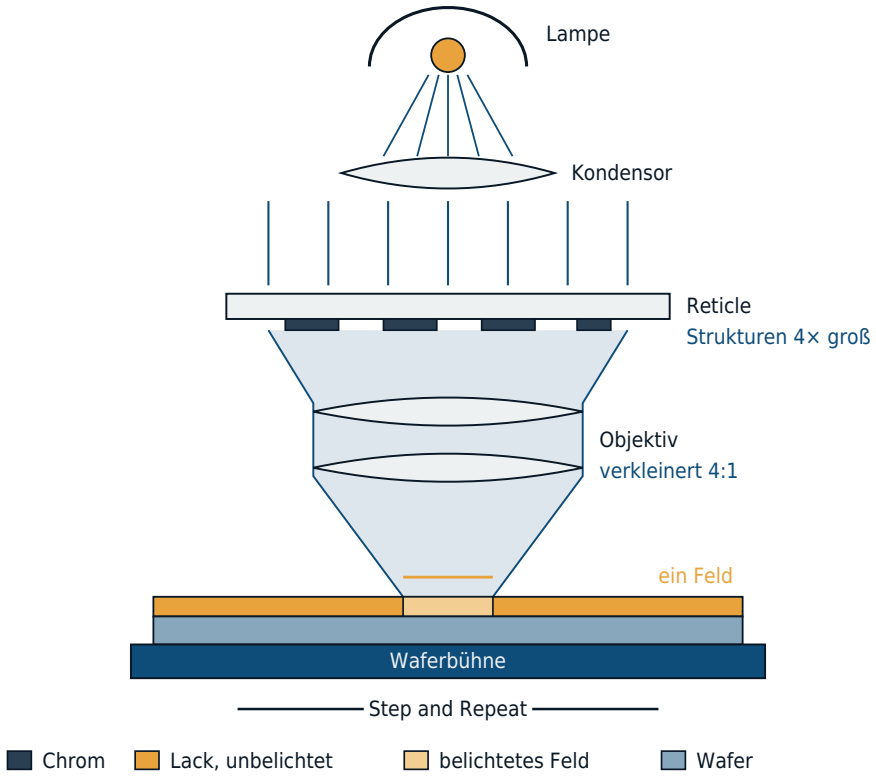
Der Spalt schützt Maske und Lack vor Kratzern und vor Partikeln dazwischen. Dafür fällt nur ein Schattenbild auf die Scheibe: Das Licht wird an den Chromkanten gebeugt, die Strukturen werden breiter und ihre Ränder unscharf.

Reduzierende Projektionsbelichtung

Dies ist das einzige Belichtungsverfahren, das in der Chipfertigung noch eingesetzt wird. Bei diesem Verfahren ist die Step-and-Repeat-Technik gebräuchlich. Dabei wird ein einzelner Chip - bei geringer Größe auch mehrere - über ein Reticle auf dem Wafer abgebildet. So wird die gesamte Fläche der Scheibe Chip für Chip belichtet.

Reduzierende Projektionsbelichtung

Das Objektiv bildet das Reticle verkleinert auf den Wafer ab



Weil die Strukturen auf dem Reticle vier- bis zehnfach größer sind, werden auch Partikel und Maskenfehler mitverkleinert und fallen oft unter die Auflösungsgrenze. Belichtet wird Feld für Feld, dazwischen verfährt die Waferbühne.

Bei hoher Auflösung wird das Feld nicht mehr als Ganzes belichtet, sondern Maske und Wafer werden gegenläufig synchron unter einem schmalen Lichtspalt hindurchgeführt (step and scan). Nur so lässt sich das Objektiv über das ganze Feld korrigiert halten. Der Vorteil bei diesem Verfahren ist, dass die auf dem Reticle abgebildeten Strukturen um den Faktor 4 vergrößert vorliegen (bei den High-NA-EUV-Anlagen in einer Achse um den Faktor 8). Bei der verkleinerten Abbildung auf den Wafer werden neben den Strukturen auch alle Fehler, wie Partikel, verkleinert dargestellt oder fallen sogar unter die Auflösungsgrenze; die Auflösung selbst wird bei diesem Verfahren verbessert.

Da für eine Ebene nicht die komplette Maske genutzt wird, können mehrere Ebenen auf einer Glasplatte aufgebracht werden, so dass die Maskenkosten gesenkt werden.

Zusätzlich kann ein Pellicle, eine dünne Folie mit Alurahmen, auf die Maske geklebt werden, wodurch Partikel von der Maske ferngehalten werden. Diese befinden sich nun außerhalb des Schärfebereichs und werden nicht abgebildet.

Daneben gibt es noch die Spiegelprojektionsbelichtung (1:1), bei der der Wafer

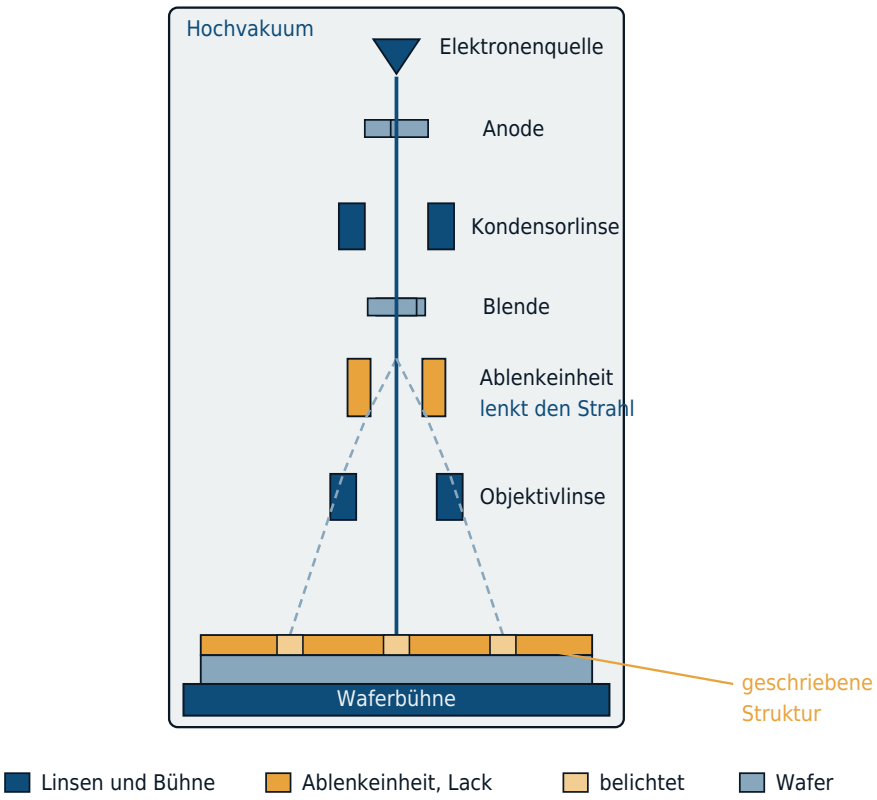
über ein komplexes System aus Spiegeln belichtet wird. Durch die Verwendung von Spiegeln entstehen keine Farbfehler, wie sie bei Linsensystem auftreten, dazu können Ausdehnungen der Scheibe durch Temperatureinfluss in Prozessen ausgeglichen werden. Jedoch werden über Spiegelsysteme Bilder verzehrt/gekrümmt dargestellt. Durch die 1:1-Abbildung ist die Auflösung außerdem stark begrenzt.

Elektronenstrahllithografie

Wie bei der Maskenherstellung wird ein fokussierter Elektronenstrahl über den belackten Wafer gescannt. Dabei kann das Scannen zeilenweise im Raster-Scan-Verfahren oder im Vektorscanverfahren durchgeführt werden. Jedoch muss auch hier jede Struktur einzeln geschrieben werden, was sehr zeitintensiv ist. Der Vorteil ist, dass keine Masken benötigt werden, was Kosten spart. Die gesamte Apparatur befindet sich dabei im Hochvakuum.

Elektronenstrahllithografie

Ein fokussierter Strahl schreibt die Struktur direkt – ohne Maske



Der Strahl fährt die Struktur Punkt für Punkt ab – zeilenweise im Rasterscan oder gezielt im Vektorscan. Das kostet Zeit, spart aber die Maske und ist deshalb vor allem für Prototypen und für die Maskenherstellung selbst üblich.

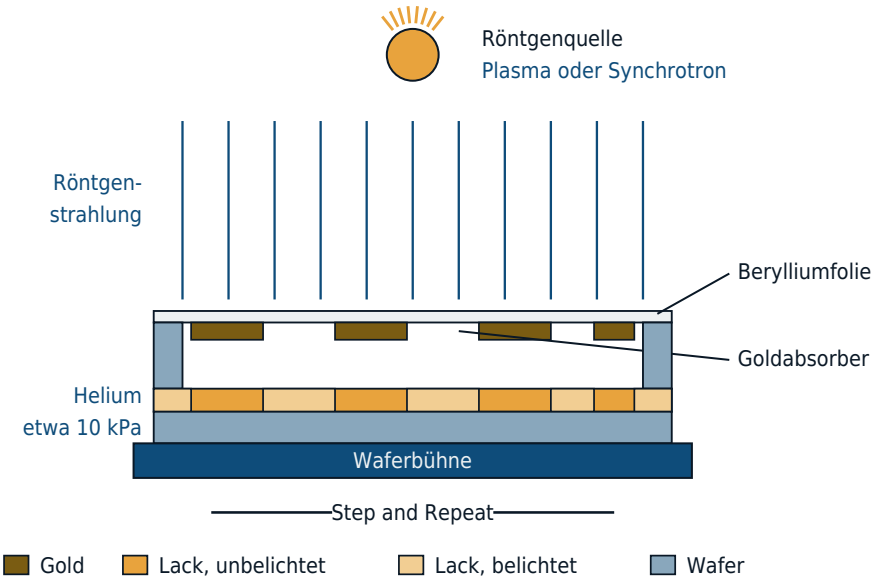
Röntgenstrahllithografie

Die Auflösungsgrenze der Röntgenstrahllithografie liegt bei etwa 40 nm. Die Abbildung erfolgt im 1:1 Step-and-Repeat-Verfahren durch Schattenwurf, und wird unter Atmosphärendruck in Luft oder bei leichtem Unterdruck in Heliumatmosphäre (ca. 10.000 Pa) durchgeführt. Die Röntgenquelle kann dabei eine Plasmaquelle oder Synchrotonstrahlung sein.

Anstelle der chrombeschichteten Glasmasken werden dünne, mechanisch stabile Folien aus Beryllium, teils auch aus Silicium verwendet. Um die Röntgenstrahlung zu absorbieren werden die Folien mit schweren Elementen wie Gold beschichtet. Die Anlagen sowie die Masken sind sehr teuer.

Röntgenstrahllithografie

Schattenwurf im Maßstab 1:1 – für Röntgenstrahlen gibt es keine Linsen



Röntgenstrahlen lassen sich nicht mit Linsen bündeln, deshalb wird das Muster direkt als Schatten übertragen – eins zu eins, Feld für Feld. Die Auflösung reicht bis etwa 40 nm, doch Anlagen und Masken sind sehr teuer.

Weitere Verfahren

Eine weitere Möglichkeit der Lithografie ist die Bestrahlung der Wafer mit Ionen. Mit den Ionen kann der Wafer sowohl über eine Maske strukturiert, als auch direkt wie bei der Elektronenstrahlmethode beschrieben werden. Im Falle von Wasserstoffionen beträgt die Wellenlänge 0,0001 nm. Mit anderen Elementen ist auch eine direkte Dotierung ohne Maskierung denkbar.

Der Fotolack

Lacktechnik

Es gibt Positiv- und Negativlacke, die bei unterschiedlichen Anwendungen eingesetzt werden. Während beim Positivlack die belichteten Stellen löslich werden, werden bestrahlte Bereiche beim Negativlack für die Entwicklerlösung unlöslich.

Charakteristik der Positivlacke:

- sehr gute Auflösung
- stabil in der Entwicklerlösung
- in Laugenlösungen entwickelbar
- nur bedingt resistent gegen Ätzung und Implantationsprozesse
- schlechte Haftung auf dem Wafer

Charakteristik der Negativlacke:

- sehr empfindlich (exaktere Strukturierung bei der Belichtung möglich als Positivlack)
- gute Haftung
- resistent gegen Ätzverfahren und Implantation
- billiger als Positivlack
- geringe Auflösung
- Xylol-Entwickler (giftig)

Zur Strukturierung der Wafer in der Fertigung werden fast ausschließlich Positivlacke verwendet. Negativlacke finden meist als Passivierungsschicht (Fotoimidschicht) Anwendung, die mittels UV-Licht ausgehärtet werden können. Sofern im Text keine Angaben zum Lack gemacht werden, handelt es sich um Positivlack; bei Negativlack wird dies ausdrücklich erwähnt.

Chemische Zusammensetzung

Fotolacke (auch als Fotoresist bezeichnet; resist = widerstehen) setzen sich aus einem Bindemittel, einem Sensibilisator und einem Lösemittel (Verdünner) zusammen.

Bindemittel (Anteil ~20%)
: Als Bindemittel werden häufig Novolake eingesetzt. Dabei handelt es sich um Phenolharze (Kunstharz, Kunststoff), welche in erster Linie die thermischen Eigenschaften des Lackes bestimmen.

Sensibilisator (Anteil ~10%)
: Der Sensibilisator bestimmt die Lichtempfindlichkeit des Lackes.

Sensibilisatoren setzen sich aus Molekülen zusammen, die bei einer Belichtung mit energiereicher Strahlung die Löslichkeit des Lackes verändern (im Positivlack entsteht aus dem Sensibilisator durch die Belichtung Carbonsäure. Mehr dazu im Kapitel Entwickeln). Damit der Lack nicht durch das Licht in den Fertigungshallen belichtet wird, finden die Fototechnikprozesse unter Gelblicht statt, gegen das der Lack unempfindlich ist.

Lösemittel (Anteil ~70%)
: Die Lösemittel bestimmen die Viskosität des Lackes. Durch das Verdampfen der Lösemittel auf einer Heizplatte wird der Lack stabilisiert und resistent für nachfolgende Prozesse.

Ein vom Hersteller gelieferter Lack hat eine definierte Oberflächenspannung und Dichte, einen bestimmten Feststoffgehalt und eine bestimmte Viskosität. Somit ist bei der Belackung in der Chipfertigung die erzielte Fotolackdicke ausschließlich von der Drehzahl der Belackungsanlage abhängig.

Chemisch verstärkte Lacke

Der oben beschriebene Aufbau gilt für die klassischen Lacke der i-Linie. Bei 248 nm und darunter reicht die Zahl der eintreffenden Photonen nicht mehr aus, um genügend Sensibilisatormoleküle direkt umzusetzen. Eingesetzt werden deshalb chemisch verstärkte Lacke (chemically amplified resist, CAR): Statt eines Sensibilisators enthalten sie einen Photosäurebildner (photo acid generator, PAG), der bei Belichtung eine starke Säure freisetzt. Diese spaltet in einem anschließenden Temperschritt (post exposure bake) Schutzgruppen vom Bindemittel ab und wird dabei nicht verbraucht, sondern wirkt katalytisch weiter. Ein einzelnes Photon löst so hunderte bis tausende Reaktionen aus.

Der Verstärkungsmechanismus hat einen Preis: Die Säure diffundiert während des Temperschritts, was die Kante der späteren Lackstruktur verschmiert. Diffusionslänge, Bake-Temperatur und -Zeit sind damit auflösungsbestimmende Prozessparameter.

Lacke für die EUV-Belichtung

Ein EUV-Photon trägt mit 92 eV rund vierzehnmal mehr Energie als ein Photon bei 193 nm. Für dieselbe eingebrachte Energie treffen also entsprechend weniger Photonen auf die Fläche, und deren Ankunft ist ein statistischer Prozess. Bei Strukturen von wenigen Nanometern schwankt die Photonenzahl pro Struktur merklich – die Folge sind zufällig fehlende oder zusätzliche Strukturen sowie eine rauere Kante. Diese stochastischen Defekte, nicht die Optik, sind heute die praktische Auflösungsgrenze der EUV-Lithografie.

Gegenmaßnahmen sind höhere Belichtungsdosen (die den Durchsatz kosten)

und Lacke mit höherer Absorption. Neben weiterentwickelten CAR kommen dafür Metalloxidlacke auf Zinnbasis in Betracht, die EUV deutlich stärker absorbieren als organische Systeme. Ebenfalls in Entwicklung sind trocken aufgebrachte Lacke, die aus der Gasphase abgeschieden statt aufgeschleudert werden.

Die Schichtdicken sind dabei deutlich geringer als in der klassischen Fototechnik: Bei EUV liegen sie im Bereich einiger zehn Nanometer, da hohe schmale Lackstege sonst durch die Oberflächenspannung der Entwicklerlösung umkippen (pattern collapse). Die mechanische Stabilität und die Ätzresistenz übernehmen deshalb zusätzliche Hilfsschichten unter dem Lack.

Entwickeln und Kontrolle

Entwickeln

Die belichteten Scheiben werden in Tauch- oder Sprühentwicklungen prozessiert. Während beim Tauchätzschritt eine komplette Horde in einer Laugenlösung entwickelt und anschließend gespült werden kann, wird bei der Sprühentwicklung jeder Wafer einzeln prozessiert. Wie beim Belacken befindet sich der Wafer auf einem Chuck und wird bei langsamer Umdrehung stetig mit Entwicklerlösung besprüht. Nach dem vollständigen Entwickeln des Lackes wird der Wafer mit aufgespritztem Wasser gespült um den Entwicklungsprozess zu stoppen. Einige Vorteile der Sprühentwicklung gegenüber dem Tauchverfahren sind folgende:

- kleinste Strukturen können freigelegt werden
- Die Entwicklerlösung wird stetig erneuert: Verunreinigungen werden verhindert
- Die Menge der eingesetzten Entwicklerlösung ist wesentlich geringer

Durch das Entwickeln lösen sich, entsprechend der Lackart, bestimmte Bereiche des Lackes, so dass am Ende eine strukturierte Scheibe bestehen bleibt. Durch die Belichtung wird im Lack eine Reaktion ausgelöst, bei der der Sensibilisator in Säure umgewandelt wird. Diese Carbonsäure wird beim Entwickeln nach folgender Gleichung mit Natronlauge (NaOH) zu wasserlöslichem Salz umgewandelt:

$R\text{-}COOH + NaOH \rightarrow (R\text{-}COO)^- + Na^+ + H_2O$ (R = weitere unbeteiligte Stoffe)

Darstellung der belichteten Lackarten nach dem Entwickeln

Da bei Kali- oder Natronlösungen Rückstände des Entwicklers auf dem Wafer zurückbleiben, werden auch Metallionen freie Entwickler wie TMAH (Tetramethylammoniumhydroxid) eingesetzt. Durch einen erneuten Aushärteschritt (Hard-Bake) wird der Lack für nachfolgende Prozesse, wie Ätzen oder für eine Ionenimplantation, beständig gemacht.

Lackkontrolle

Nun wird die Lackstruktur kontrolliert. Unter schrägem Lichteinfall lassen sich im Mikroskop die Gleichmäßigkeit der Lackschicht, sowie schlechte Fokussierungen und Lackanhäufungen erkennen. Sind die Lackstege zu dick oder dünn muss der Lack entfernt und der Prozess wiederholt werden. Ebenso muss die Lackstruktur zu der sich darunter befindlichen Ebene exakt justiert sein, andernfalls ist ebenso eine Wiederholung der Belackungs- und Belichtungsprozesse notwendig. Dazu gibt es unterschiedliche Justiermarken für die Justiergenauigkeit und die Linienweitenkontrolle:

Darstellung von Justiermarken

Die Breite der Lackstege wird mit einem Mikroskop kontrolliert: Lichtstrahlen treffen senkrecht auf den Wafer, und werden an Kanten nicht zum Objektiv zurückgestreut. So sind schwarze Linien als Strukturbegrenzungen sichtbar, mit deren Abständen zueinander man unter zu Hilfenahme der Mikroskopvergrößerung die Breite der Stege berechnen kann.

Lackentfernen

Nachdem die Struktur unter dem Lack in Ätzverfahren abgetragen wurde, oder der Lack beim Implantationsprozess als Maskierungsschicht gedient hat, muss der Lack entfernt werden. Dies geschieht mit starken Ätzlösungen (Remover), in einem Trockenätzschritt oder mit Lösungsmitteln. Als Lösungsmittel zum Entfernen der Lackschicht eignet sich Aceton, da es den Wafer und andere Schichten nicht angreift. Durch eine Ionenimplantation oder einen Trockenätzprozess kann die Lackschicht jedoch weiter ausgehärtet sein, so dass Lösungsmittel den Lack nicht mehr angreifen und entfernen können.

In diesem Fall kann der Lack mit einem Remover bei ca. 80 °C in einem Tauchverfahren entfernt werden. Hat sich der Lack während der Bearbeitung

auf über 200 °C erwärmt kann er auch vom Remover nicht mehr abgetragen werden. Dann muss der Lack mittels Lackveraschung oder Lackverbrennung entfernt werden.

Unter Anwesenheit von Sauerstoff wird dabei durch Hochfrequenzanregung eine Gasentladung gezündet, so dass angeregte Sauerstoffatome entstehen. Diese verbrennen den Lack rückstandsfrei. Die geladenen Teilchen werden jedoch durch das elektrische Feld stark beschleunigt und können so einen leichten Abtrag der Scheibenoberfläche oder eine geringe Schädigung der Scheibe verursachen. Das Lackveraschen (Strippen) kann direkt nach dem Ätzen noch in derselben Prozesskammer (in-situ) oder in einer anderen Kammer (ex-situ) erfolgen.

Maskentechnik

Maskentechnik

Die in der Fototechnik eingesetzten Masken enthalten ein Muster mit dem die jeweilige Schicht auf dem Wafer strukturiert wird. Ausgangsmaterial für die Masken sind Glasplatten, die ganzflächig mit Chrom und Lack beschichtet sind (Blanks). Über den für Elektronenstrahlen empfindlichen Lack wird die Chromschicht strukturiert, welche dann die lichtundurchlässigen Bereiche auf der Glasmaske darstellt.

Mit einem Elektronenstrahl werden die Masken direkt beschrieben. Die gesamte Apparatur – die Elektronenstrahlquelle, Fokussier- und Ablenkeinheit und der Blank – befindet sich im Hochvakuum (0,01–100 Pa; normaler Luftdruck ca. 100.000 Pa). Der Elektronenstrahl wird computergesteuert über die Maske gelenkt und belichtet den Lack. Bei dieser Methode lassen sich Strukturen bis weit unter 100 nm auflösen.

Ein einzelner Strahl schreibt eine hochaufgelöste Maske allerdings nicht in vertretbarer Zeit. Heutige Maskenschreiber arbeiten deshalb mit mehreren hunderttausend parallel geführten Einzelstrahlen (multi beam mask writer), die gemeinsam über den Blank geführt und einzeln ein- und ausgeschaltet werden. Erst damit sind Schreibzeiten von einigen Stunden pro Maske erreichbar – und erst damit lassen sich beliebig geformte, gekrümmte Strukturen genauso schnell schreiben wie rechtwinklige.

Korrektur der Abbildungsfehler

Auf Grund wellenoptischer Effekte (z.B. Beugung) kann es bei der Belichtung

der Wafer zu Abbildungsfehlern kommen, welche durch die sogenannte *optical proximity correction* (OPC, etwa: optische Nahbereichskorrektur) korrigiert bzw. verringert werden.

Dazu können die eigentlichen Strukturen so abgeändert werden, dass die Abbildung auf dem Wafer dem gewünschten Muster entspricht. Außerdem können zusätzliche Hilfsstrukturen auf die Maske geschrieben werden welche nur dazu dienen, Abbildungsfehler zu minimieren, jedoch keine Funktion für die Schaltung haben.

Bei den kleinsten Strukturen genügt es nicht mehr, die Maske allein zu korrigieren. Optimiert werden Maske und Beleuchtung gemeinsam (source mask optimization, SMO): Auch die Winkelverteilung des einfallenden Lichts wird auf das jeweilige Muster zugeschnitten. Die weitestgehende Variante ist die inverse Lithografie (inverse lithography technology, ILT). Dabei wird nicht mehr eine gezeichnete Maske korrigiert, sondern aus dem gewünschten Ergebnis auf dem Wafer rückwärts diejenige Maskenstruktur berechnet, die dieses Ergebnis erzeugt. Heraus kommen frei geformte, oft gekrümmte Konturen, die mit einer gezeichneten Vorlage keine Ähnlichkeit mehr haben. Die Rechenzeit dafür übersteigt inzwischen die Schreibzeit der Maske deutlich und wird auf großen Rechnerclustern erbracht.

Weil eine einzige Maske viele tausend Wafer belichtet, überträgt sich jeder Fehler auf ihr auf jeden Chip. Masken werden deshalb nach dem Schreiben mit hochauflösenden Verfahren geprüft und einzelne Defekte gezielt repariert – etwa durch Abtragen überschüssigen Absorbermaterials mit einem fokussierten Elektronen- oder Ionenstrahl.

Maskenherstellung

Grundsätzlich werden die Strukturen auf den Masken auf die gleiche Weise hergestellt, wie auf den Wafern. Im Gegensatz zum Belichtungsprozess in der Waferfertigung, bei dem die Strukturen der Maske durch Schattenwurf auf dem Lack abgebildet werden, werden die Masken aber mit einem Elektronenstrahl direkt beschrieben.

1. Belichten eines fotoempfindlichen Lacks zur Strukturierung einer Chromschicht auf dem Glassubstrat mittels Laser oder Elektronenstrahl

2. Entwickeln der Lackschicht

3. Ätzen der Chromschicht

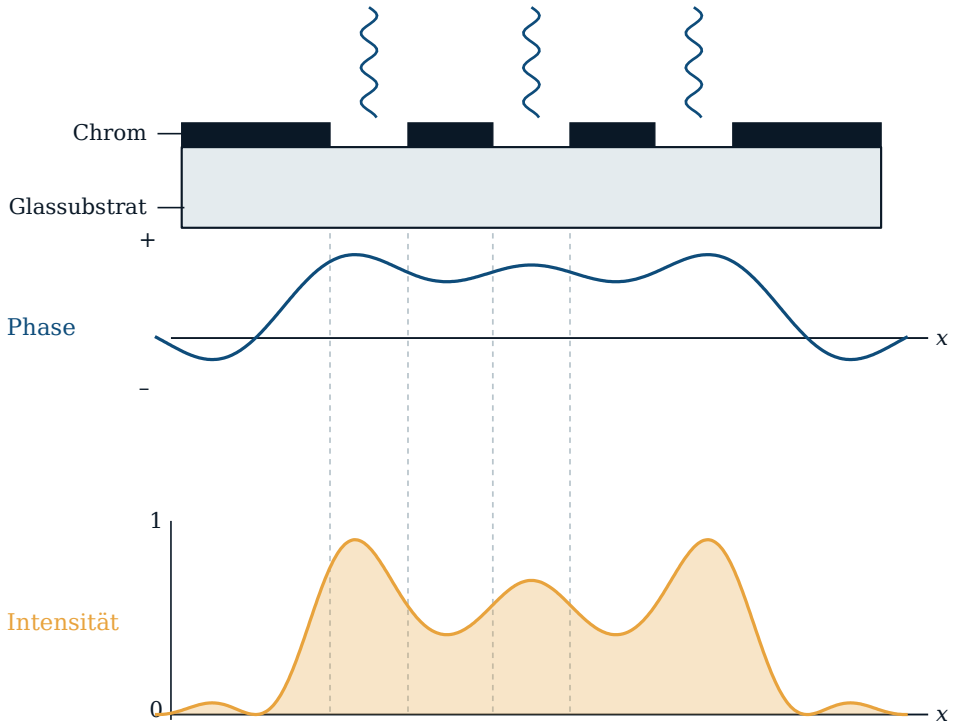
4. Lackentfernen

5. Anbringen des Pellicles

Maskentypen

Neben der klassischen Chrommaske (COG, Chrome On Glass) gibt es noch weitere Maskentypen, die eine verbesserte Strukturauflösung ermöglichen. Hauptproblem der COG-Maske ist die Beugung, die das Licht an den Strukturkanten erfährt. Dadurch fällt das Licht nicht nur senkrecht auf den Wafer, sondern wird auch in den Schattenbereich abgelenkt, wo der Fotolack nicht belichtet werden soll.

Intensitätsprofil einer Chrom On Glass-Mask

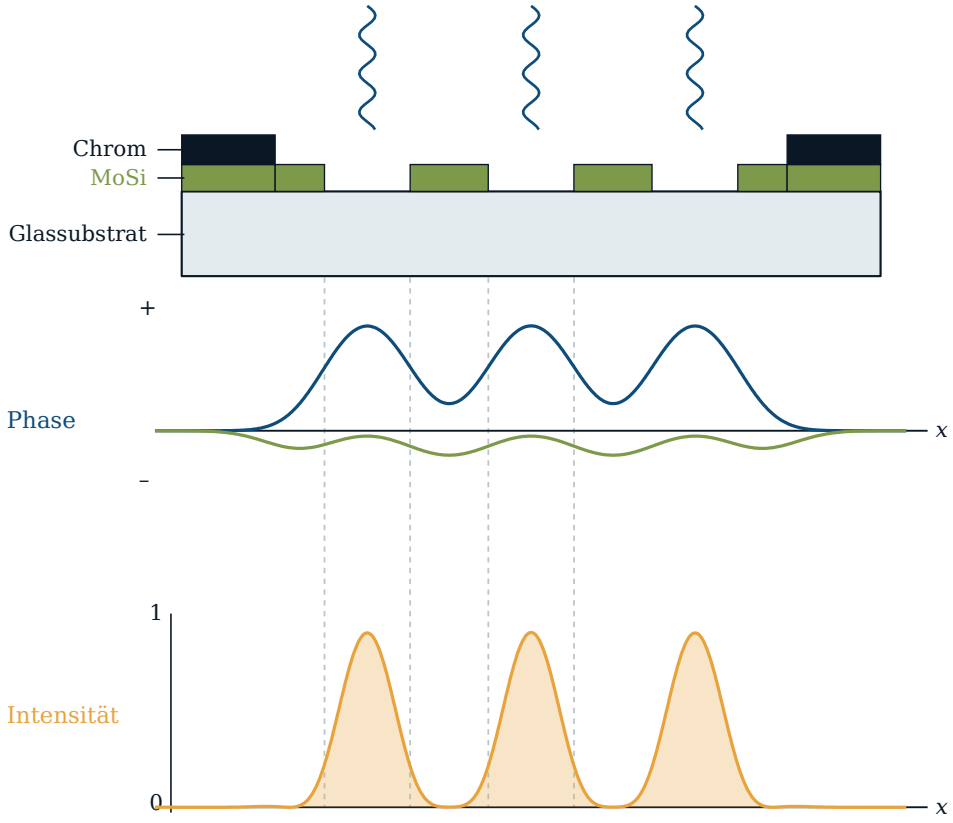


Mit verschiedenen Maßnahmen versucht man nun, die Intensität des gebeugten Lichts zu vermindern. Diese werden im Folgenden anhand der unterschiedlichen Maskentypen näher beschrieben.

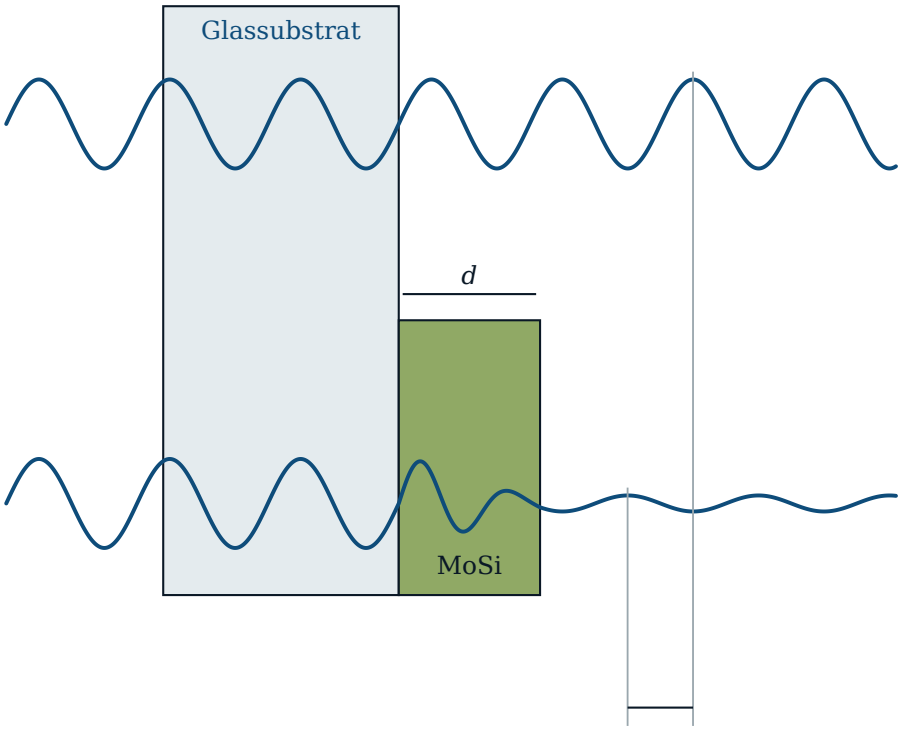
Attenuated Phase Shift Mask (AttPSM)

Bei der so genannten Halbtonmaske oder weichen Phasenmaske bildet eine Schicht aus Molybdänsilicid (MoSi) den strukturgebenden Teil, eine Chromschicht gibt es hier nicht. Die Dicke der MoSi-Schicht ist so gewählt, dass das Licht beim Durchgang eine Phasenverschiebung um 180° erfährt gegenüber dem Licht, das lediglich Glas durchläuft. Dies geschieht durch die unterschiedliche Lichtgeschwindigkeit in Luft und in MoSi. Gleichzeitig ist die Schicht je nach Molybdänanteil im Silicium zu 6 oder 18 % lichtdurchlässig (bei einer Belichtungswellenlänge von 193 nm), das Licht wird also abgeschwächt (attenuate, engl.: vermindern). Die gegenläufigen Lichtwellen löschen sich so unterhalb der MoSi-Strukturen nahezu aus, somit wird der Kontrast zwischen Hell und Dunkel erhöht. Zusätzlich kann in Bereichen, die nicht zur Belichtung benötigt werden, Chrom aufgebracht werden, um Licht vollständig auszublenden. Diese Masken werden als Tritonemaske bezeichnet.

Intensitätsprofil einer Attenuated Phase Shift Mask



Prinzip der Phasenschiebung mittels Molybdänsilicid



Phasenverschiebung um 180°

Chromfreie Phasenschiebermaske

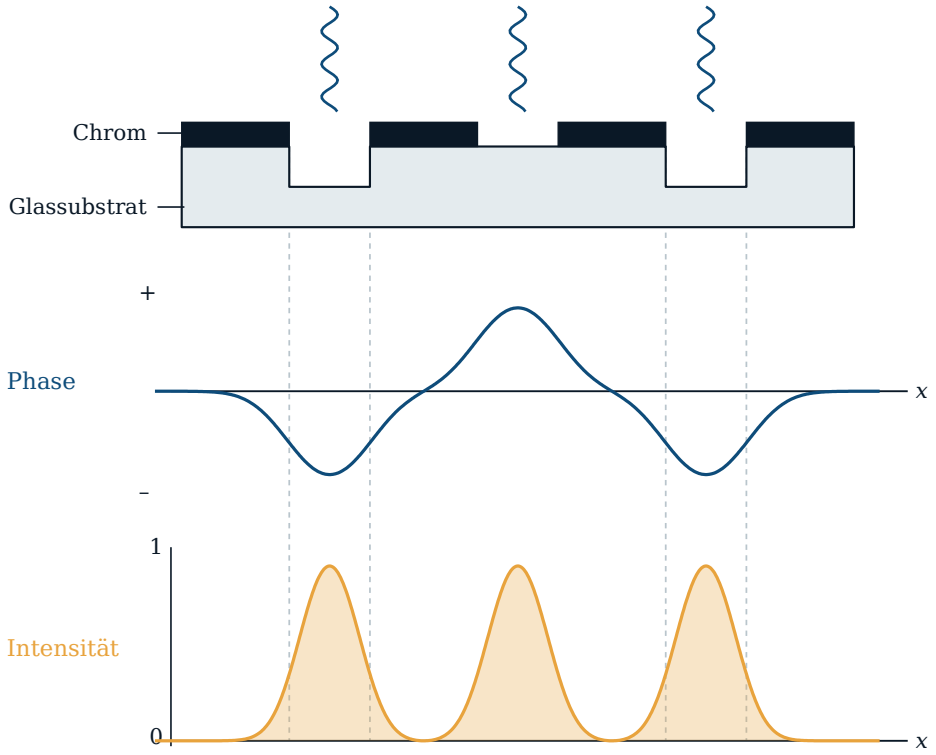
Die chromfreien Masken besitzen keine strukturgebende Beschichtung. Die Phasenverschiebung wird durch Gräben erzeugt, die direkt in die Glasplatte geätzt sind. Die Herstellung dieser Masken gestaltet sich schwierig, da der Ätzvorgang mitten im Glas gestoppt werden muss. Im Gegensatz zu Ätzprozessen, bei denen eine Schicht vollständig durchgeätzt wird und Änderungen im Plasma Auskunft darüber geben, wann die darunterliegende Schicht freigelegt ist, erhält man hier keine Information, wann die benötigte Tiefe erreicht ist.

Zusätzlich ergeben sich Probleme bei der Herstellung großer Strukturen. Dadurch, dass es keine abschattenden Bereiche gibt und das Licht überall mit derselben Intensität auf die Maske trifft, erfolgt nur an den Strukturkanten die gewünschte Interferenz und die daraus resultierende Abschwächung des Lichts. In der Mitte der Bereiche findet jedoch keine oder nur noch eine geringe Abschwächung statt. Um eine Belichtung dieser Bereiche zu verhindern, muss die Lichtintensität also von vorn herein reduziert werden, wodurch die Gefahr einer Unterbelichtung aller Strukturen entsteht.

Alternating Phase Shift Mask (AltPSM)

Bei der alternierenden Phasenmaske werden wie bei der chromfreien Maske Gräben direkt in das Glassubstrat geätzt, jedoch abwechselnd (alternierend) mit ungeätzten Bereichen. Zusätzlich werden Stellen mit Chrom beschichtet um die Lichtintensität an entsprechenden Stellen herabzusetzen.

Intensitätsprofil einer Alternating Phase Shift Mask



Dadurch ergeben sich jedoch Bereiche mit undefinierter Phasenverschiebung, so dass bei diesem Maskentyp, der eine sehr hohe Auflösung ermöglicht, grundsätzlich zweimal belichtet werden muss. Die erste Maske enthält dabei die Strukturen die in x-Richtung verlaufen, die zweite Maske die Struktur in y-Richtung.

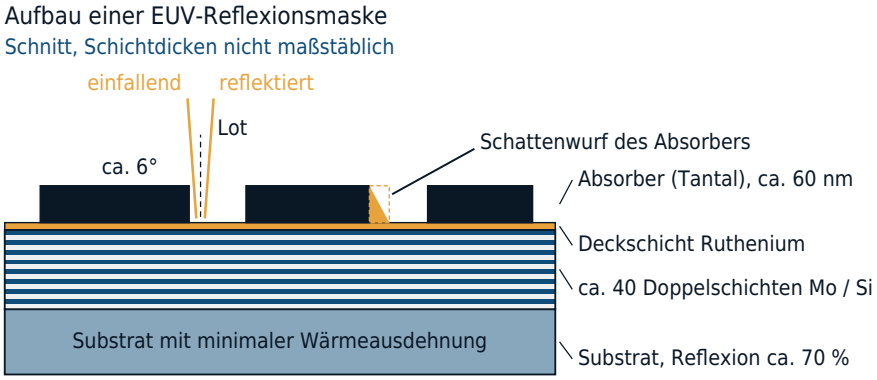
Zielstruktur auf dem Wafer und entsprechende Maskenstruktur

Reflexionsmasken für die EUV-Belichtung

Alle bisher beschriebenen Maskentypen sind Transparente: Licht tritt durch das Glassubstrat hindurch. Für EUV mit 13,5 nm funktioniert das nicht, weil jedes Material diese Strahlung absorbiert. Die EUV-Maske ist deshalb ein Spiegel.

Als Substrat dient ein Glas mit nahezu verschwindender Wärmeausdehnung, da sich die Maske unter der Bestrahlung sonst verzieht. Darauf liegen etwa vierzig Doppelschichten aus Molybdän und Silicium, deren Dicken so gewählt sind, dass sich die an den einzelnen Grenzflächen reflektierten Anteile konstruktiv überlagern. Selbst dieser Vielschichtspiegel reflektiert nur rund 70 % der auftreffenden Strahlung; da im Strahlengang mehrere solche Spiegel liegen, erreicht nur ein kleiner Teil der erzeugten Leistung den Wafer. Eine dünne Deckschicht aus Ruthenium schützt den Stapel, darüber bildet ein Absorber, meist eine Tantalverbindung, das eigentliche Muster.

Aufbau einer EUV-Reflexionsmaske



Aus dem Spiegelprinzip folgt eine Eigenheit, die es bei Transparentmasken nicht gibt: Der Strahl muss schräg einfallen, damit einfallendes und reflektiertes Licht getrennt werden können. Bei diesem Einfall von einigen Grad wirft der etwa 60 nm dicke Absorber einen Schatten, und zwar je nach Orientierung der Struktur einen unterschiedlichen. Diese Maskeneffekte müssen bei der Korrektur der Maske mitgerechnet werden.

Auch der Schutz vor Partikeln ist schwieriger. Ein Pellicle muss hier von der Strahlung zweimal durchlaufen werden und darf sie dabei kaum absorbieren; gleichzeitig muss die Membran der Wärmelast einer Hochleistungsquelle standhalten. Verwendet werden dafür extrem dünne Membranen, unter anderem auf Basis von Kohlenstoffnanoröhren.

Phasenschiebende EUV-Masken, die den Kontrast wie bei 193 nm zusätzlich erhöhen würden, sind Gegenstand der Entwicklung, in der Fertigung aber noch

nicht etabliert. Die Auflösung wird bei EUV daher vor allem über die numerische Apertur, die Beleuchtung und den Lack gewonnen, nicht über die Maskentechnik.

Belichtungsverfahren der nächsten Generation

Der Wechsel, den dieses Kapitel jahrelang als bevorstehend beschrieben hat, ist vollzogen: Die EUV-Lithografie mit 13,5 nm ist seit 2019 in der Serienfertigung und belichtet die kritischen Ebenen aller führenden Logik- und Speicherprozesse. Sie arbeitet, wie vorhergesagt, im Vakuum, bildet über Spiegel ab und nutzt Masken mit spiegelnder statt transparenter Oberfläche.

Gleichzeitig ist die 193-nm-Immersionsbelichtung nicht verschwunden. Sie belichtet weiterhin die Mehrzahl der Ebenen eines modernen Bauelements, weil sie schneller und je Belichtung deutlich günstiger ist. EUV wird nur dort eingesetzt, wo es sich rechnet: Wo eine EUV-Belichtung zwei oder mehr Immersionsschritte samt Zwischenprozessen ersetzt, ist sie im Vorteil.

High-NA-EUV

Die aktuelle Ausbaustufe erhöht die numerische Apertur von 0,33 auf 0,55 und erreicht damit rund 8 nm Halbperiode in einer Belichtung. Erkauft wird das mit einer anamorphen Optik, die in den beiden Achsen unterschiedlich stark verkleinert und damit nur noch das halbe Feld belichtet, sowie mit erheblichem Aufwand: Eine Anlage kostet etwa das Doppelte einer herkömmlichen EUV-Anlage und braucht am Standort Monate bis zur Qualifikation. Erste Ebenen eines Serienprodukts wurden im Juli 2026 freigegeben, der breite Einsatz ist für 2027 und 2028 zu erwarten. Nicht alle Hersteller gehen diesen Weg – teils wird die 0,33er-Generation mit Mehrfachbelichtung als wirtschaftlicher eingeschätzt.

Wo die Grenze inzwischen liegt

Die optische Auflösung ist nicht mehr der eigentliche Engpass. Begrenzend wirken:

- Statistik: die stochastischen Defekte durch die geringe Photonenzahl pro Struktur (siehe Fotolack)
- Überlagerung: der zulässige Fehler bei der Justierung mehrerer Ebenen zueinander liegt im Bereich weniger Nanometer und wird bei geteilten Feldern und Mehrfachbelichtung weiter aufgezehrt
- Lichtleistung: der Durchsatz einer EUV-Anlage hängt direkt an der Leistung der Zinn-Plasmaquelle und an der nötigen Belichtungsdosis
- Kosten: Anlagen und Masken sind so teuer, dass sich feinste Strukturen nur bei sehr hohen Stückzahlen rechnen

Alternative Ansätze

Neben der weiteren Steigerung der Apertur werden Verfahren verfolgt, die ohne abbildende Optik arbeiten:

Nanoimprint
: Eine Vorlage mit dem Negativ der Struktur wird in eine flüssige Lackschicht gepresst, die anschließend aushärtet. Ohne Beugungsgrenze und ohne teure Optik, dafür mit den Problemen jedes Kontaktverfahrens: Defekte und Verschleiß der Vorlage. Im Einsatz für Speicherbauelemente und optische Komponenten.

Direktschreiben mit Elektronen
: Vielstrahlsysteme schreiben die Struktur maskenlos direkt auf den Wafer. Für Prototypen, Kleinserien und kundenspezifische Bausteine attraktiv, für die Massenfertigung nach wie vor zu langsam.

Gerichtete Selbstorganisation
: Blockcopolymere ordnen sich innerhalb einer lithografisch grob vorgegebenen Führungsstruktur selbst zu regelmäßigen feinen Mustern. Geeignet für streng periodische Strukturen wie Kontaktlochraster, nicht für beliebige Layouts.

Keiner dieser Ansätze ersetzt die optische Projektion in der Logikfertigung. Die Entwicklung der vergangenen zwanzig Jahre hat vielmehr gezeigt, dass ein etabliertes Belichtungsverfahren durch Maskentechnik, Beleuchtung, Rechenkorrektur und Mehrfachbelichtung sehr viel weiter getragen werden kann, als es die Wellenlänge zunächst erwarten lässt.

Optical Proximity Correction (OPC)

Einführung & Motivation

Wenn die Maske nicht mehr das druckt, was sie zeigt

Eine Fotomaske enthält im Idealfall exakt die Geometrie, die später auf dem Wafer erscheinen soll: ein Rechteck auf der Maske ergibt ein Rechteck im Resist. Dieses Ideal gilt jedoch nur, solange die kleinsten Strukturen auf der Maske deutlich größer sind als die Wellenlänge des verwendeten Lichts. Sobald Strukturgröße und Wellenlänge in dieselbe Größenordnung geraten, dominieren Beugungseffekte das Abbildungsverhalten – Kanten werden nicht mehr scharf abgebildet, sondern durch das optische System systematisch verzerrt.

Dieses Verhältnis wird durch den sogenannten k1-Faktor beschrieben, der die erreichbare Auflösung ins Verhältnis zu Wellenlänge und Apertur des

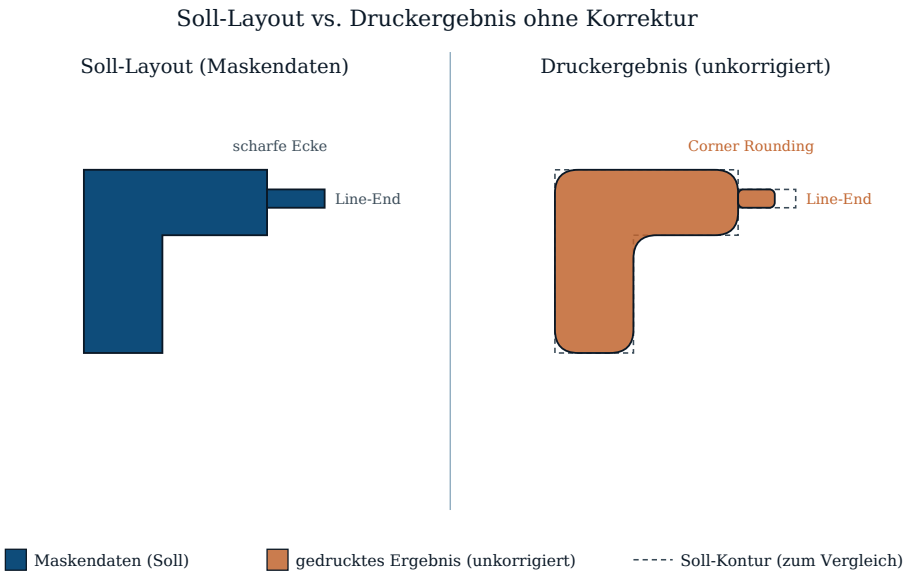
Belichtungssystems setzt. Moderne Prozesse arbeiten mit k1-Werten deutlich unterhalb dessen, was klassische, unkorrigierte Optik noch scharf abbilden kann – die Wellenlänge der Belichtung (in vielen Nodes weiterhin 193 nm) ist damit größer als die zu druckende Strukturbreite selbst. Bei einem derart „unterauflösenden" Prozess reicht eine reine 1:1-Übertragung der gewünschten Geometrie nicht mehr aus.

Optical Proximity Correction: die Maske vorverzerren

Optical Proximity Correction (OPC) löst dieses Problem, indem sie das Problem umkehrt: Statt zu versuchen, die Optik perfekter zu machen, wird die Maskengeometrie so vorverzerrt, dass die bekannten, vorhersagbaren Abbildungsfehler des optischen Systems das gewünschte Ergebnis auf dem Wafer erzeugen. Eine Ecke, die ohne Korrektur abgerundet würde, wird auf der Maske bereits überzeichnet, damit sie nach der Abbildung wieder scharf erscheint.

OPC ist damit kein nachträglicher Reparaturschritt, sondern fester Bestandteil des Maskendatenflusses: Jedes moderne Layout durchläuft vor der eigentlichen Maskenherstellung eine automatisierte OPC-Software, die auf Basis eines kalibrierten Modells des Belichtungsprozesses die Geometrie chipweit anpasst. Ohne diesen Schritt wären die meisten heutigen Strukturgrößen mit der verfügbaren Belichtungswellenlänge schlicht nicht fertigbar.

Das folgende Diagramm zeigt das Grundproblem an einem einfachen Beispiel: links das gewünschte Layout, rechts das Ergebnis, das ohne Korrektur tatsächlich gedruckt würde.



Typische Druckfehler ohne Korrektur

Corner Rounding und Line-End Shortening

Die im vorigen Abschnitt gezeigten Effekte sind die unmittelbarste Folge der Beugung: Scharfe, rechtwinklige Ecken einer Maskenstruktur werden im Resist zu abgerundeten Konturen, weil das optische System hochfrequente Ortsfrequenzanteile – die für scharfe Kanten nötig wären – nicht mehr überträgt. Bei konvexen Ecken „frisst" sich die Rundung in die Struktur hinein, bei konkaven Ecken (etwa in einer inneren Ecke eines L-förmigen Layouts) füllt sie die Kerbe stattdessen auf.

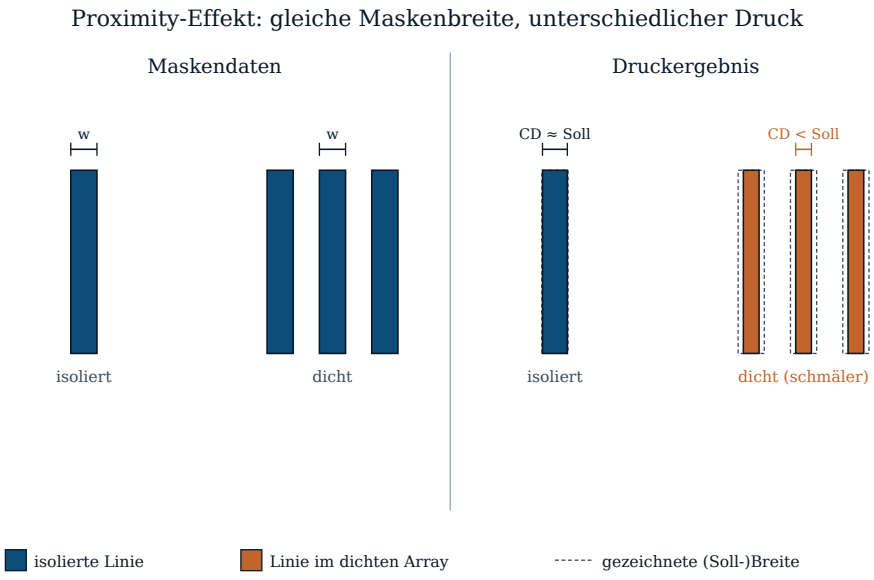
Am Ende einer Leiterbahn tritt ein verwandtes Phänomen auf: das Line-End Shortening. Weil das optische System das Ende einer Linie nicht als scharfe Kante, sondern als allmählich abfallende Intensität abbildet, zieht sich der tatsächlich entwickelte Resist am Leitungsende gegenüber der Maskenvorgabe zurück – eine Leitung, die exakt bis zu einem bestimmten Kontakt reichen sollte, endet im Druck möglicherweise einige Nanometer davor.

Der Proximity-Effekt: gleiche Breite, unterschiedliches Ergebnis

Der namensgebende Effekt von Optical Proximity Correction ist jedoch ein dritter, subtilerer: Zwei Linien mit exakt derselben auf der Maske gezeichneten Breite drucken unterschiedlich breit, je nachdem, wie viele weitere Strukturen sich in ihrer unmittelbaren Nachbarschaft befinden. Eine isoliert stehende Linie erhält ihr Beugungsbild ausschließlich von den eigenen Kanten. Eine Linie inmitten eines dichten Linienarrays dagegen überlagert sich mit den Beugungsmustern ihrer Nachbarn – das resultierende Intensitätsprofil im Resist unterscheidet sich systematisch von dem der isolierten Linie, obwohl die Maskengeometrie identisch ist.

Dieser sogenannte Iso-Dense-Bias ist über weite Bereiche des Pitch-Spektrums (Abstand von Strukturmitte zu Strukturmitte) hinweg unterschiedlich stark ausgeprägt und verläuft nicht linear – er muss für jeden Prozess separat charakterisiert werden, üblicherweise durch systematisches Belichten und Vermessen von Teststrukturen mit variierendem Pitch. Das Ergebnis dieser Charakterisierung fließt direkt in das Korrekturmodell ein, das im weiteren Verlauf dieses Kapitels vorgestellt wird.

Das folgende Diagramm zeigt den Proximity-Effekt an einer isolierten Linie im Vergleich zu einer Linie inmitten eines dichten Arrays – beide mit identischer Maskenbreite, aber unterschiedlichem Druckergebnis.



Mask Bias & Serifs

Mask Bias: die Maßkorrektur

Die einfachste Form der Korrektur setzt direkt am Proximity-Effekt aus dem vorigen Abschnitt an: Ist bekannt, dass eine Linie in einem dichten Array systematisch schmaler druckt als vorgesehen, wird sie auf der Maske einfach von vornherein etwas breiter gezeichnet – der sogenannte Mask Bias. Da der Betrag dieser Abweichung vom lokalen Pitch abhängt, ist der Bias keine feste Konstante, sondern wird anhand der zuvor charakterisierten Iso-Dense-Kurve für jede Umgebungssituation im Layout separat bestimmt: dicht gepackte Bereiche erhalten einen anderen Bias-Wert als isolierte Linien oder mittlere Pitch-Bereiche.

Mask Bias korrigiert damit ausschließlich die Linienbreite als Ganzes – er verändert nicht die Form einer Ecke oder eines Leitungsendes. Für diese lokalen, geometrisch begrenzten Fehler braucht es einen zweiten, gezielteren Korrekturmechanismus.

Serifs und Hammerheads: gezielte Ecken-Verstärkung

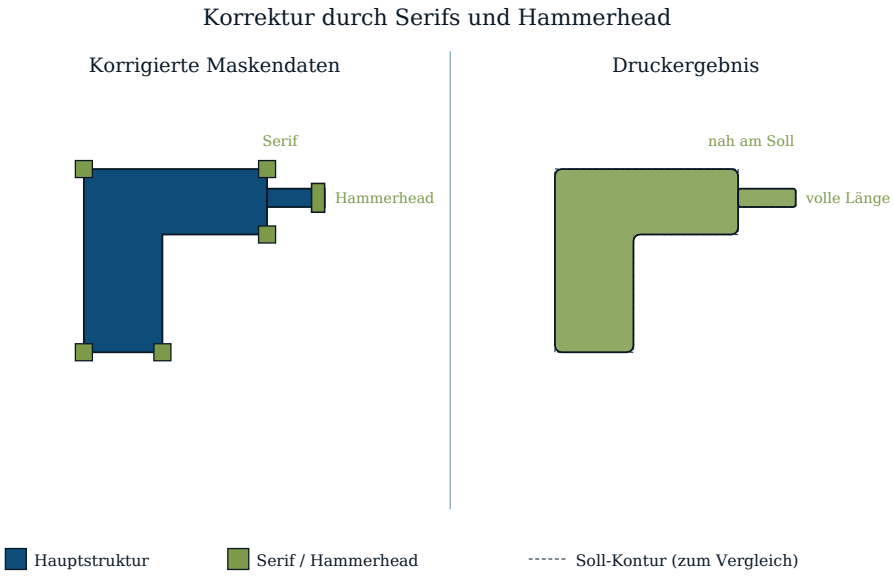
Um Corner Rounding entgegenzuwirken, wird an besonders empfindlichen Ecken – vor allem an konvexen Außenecken – eine kleine zusätzliche quadratische Struktur direkt auf der Maske ergänzt, ein sogenannter Serif. Der Serif selbst wird ebenfalls unterhalb der Auflösungsgrenze bleiben und nicht als eigenständige Struktur gedruckt; er verändert jedoch lokal die Lichtintensität am Rand der Hauptstruktur so, dass die Ecke nach der Abbildung schärfer

erscheint, als sie es ohne diese Verstärkung täte.

Am Ende einer Leitung übernimmt eine vergrößerte Variante desselben Prinzips die Korrektur des Line-End Shortening: ein sogenannter Hammerhead – eine lokale Verbreiterung des Leitungsendes senkrecht zur Leitungsrichtung – kompensiert den systematischen Rückzug des Resists am Leitungsende, sodass die Leitung auch nach der optischen Abbildung noch bis zur vorgesehenen Position reicht.

Beide Techniken – Serifs und Hammerheads – lassen sich als einfache geometrische Regeln formulieren: „Füge an jeder konvexen Ecke mit Innenwinkel kleiner als X ein Quadrat der Größe Y hinzu" oder „Verbreitere jedes Leitungsende um Z nm". Diese regelbasierte Herangehensweise ist die einfachste Form von OPC und wird im übernächsten Abschnitt der modellbasierten Variante gegenübergestellt.

Das folgende Diagramm zeigt die Layoutgeometrie aus Abschnitt 1 nach Ergänzung von Serifs und Hammerhead, sowie das dadurch deutlich verbesserte Druckergebnis.



Sub-Resolution Assist Features

Das Problem: isolierte Linien verhalten sich anders

Serifs und Mask Bias korrigieren die Form und Breite einer Struktur direkt an ihrer eigenen Kontur. Für ein grundlegenderes Problem reicht das jedoch nicht aus: Eine isolierte Linie erhält ihr Beugungsbild – wie in Abschnitt 2 beschrieben – ausschließlich von den eigenen Kanten, während eine Linie in

einem dichten Array von den Beugungsmustern ihrer Nachbarn profitiert. Diese unterschiedliche Beugungssituation wirkt sich nicht nur auf die gedruckte Breite aus, sondern auch darauf, wie tolerant die jeweilige Struktur gegenüber Fokus- und Dosisschwankungen im Belichtungsprozess ist – ihr sogenanntes Prozessfenster. Isolierte Strukturen haben typischerweise ein deutlich kleineres Prozessfenster als dicht gepackte, selbst wenn beide nach Korrektur exakt dieselbe Breite drucken.

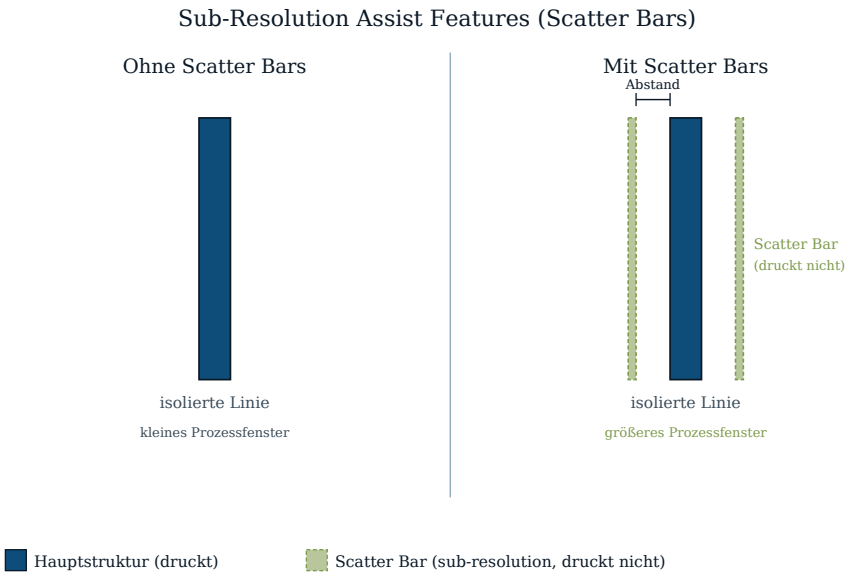
Serif und Bias allein können dieses Defizit nicht beheben, da sie nur die Struktur selbst verändern, nicht aber deren optische Umgebung.

Scatter Bars: Nachbarn, die nicht gedruckt werden

Sub-Resolution Assist Features (SRAFs), umgangssprachlich auch Scatter Bars genannt, lösen dieses Problem, indem sie der isolierten Linie künstlich Nachbarschaft verschaffen: Parallel zur eigentlichen Struktur werden zusätzliche, sehr schmale Linien auf die Maske gezeichnet – schmal genug, dass sie selbst unterhalb der Auflösungsgrenze des Belichtungssystems bleiben und daher nicht als eigenständige Struktur im Resist erscheinen. Sie verändern jedoch das Beugungsbild am Ort der Hauptstruktur so, dass es dem einer dicht gepackten Umgebung ähnlicher wird – die isolierte Linie „sieht" optisch nun fast so aus wie eine Linie im Array, mit einem entsprechend größeren, stabileren Prozessfenster.

Die Platzierung von SRAFs unterliegt strengen geometrischen Regeln: Der Abstand zur Hauptstruktur muss groß genug sein, um Verschmelzung zu vermeiden, aber klein genug, um noch einen Effekt zu haben; die Breite der Scatter Bars selbst muss sicher unterhalb der Druckschwelle bleiben, auch unter ungünstigen Prozessbedingungen. Softwaregestützte SRAF-Platzierung generiert diese Hilfsstrukturen heute automatisch anhand des charakterisierten Beugungsverhaltens des jeweiligen Prozesses, typischerweise als letzter Schritt vor der eigentlichen Serif- und Bias-Korrektur.

Das folgende Diagramm zeigt eine isolierte Linie ohne und mit Scatter Bars – die Hilfsstrukturen selbst erscheinen nicht im gedruckten Resist.



Rule-based vs. Model-based OPC

Regelbasierte OPC: schnell, aber grobmaschig

Die in den vorigen Abschnitten vorgestellten Korrekturen – Mask Bias, Serifs, Scatter Bars – lassen sich als Nachschlagetabelle organisieren: Für eine gegebene Linienbreite und einen gegebenen Abstand zur nächsten Nachbarstruktur liefert die Tabelle einen vorab charakterisierten Korrekturwert. Diese regelbasierte OPC ist rechnerisch günstig, da für jede Kante nur ein einziger Tabellen-Lookup nötig ist, keine aufwendige Simulation.

Die Grenzen dieses Ansatzes zeigen sich bei komplexer, zweidimensionaler Geometrie: In der Nähe einer Ecke, eines Line-Ends oder einer ungewöhnlichen Nachbarschaftskonstellation überlagern sich mehrere Proximity-Effekte gleichzeitig, für die keine einzelne Tabellenzeile mehr eine passende Antwort liefert. Regelbasierte OPC behandelt eine Kante dabei stets als Ganzes: Der gesamte gerade Kantenabschnitt wird um denselben Betrag verschoben, unabhängig davon, ob sich in der Nähe des einen Endes eine zusätzliche Struktur befindet, die dort einen anderen Korrekturbedarf erzeugt als am anderen Ende.

Modellbasierte OPC: Segment für Segment simuliert

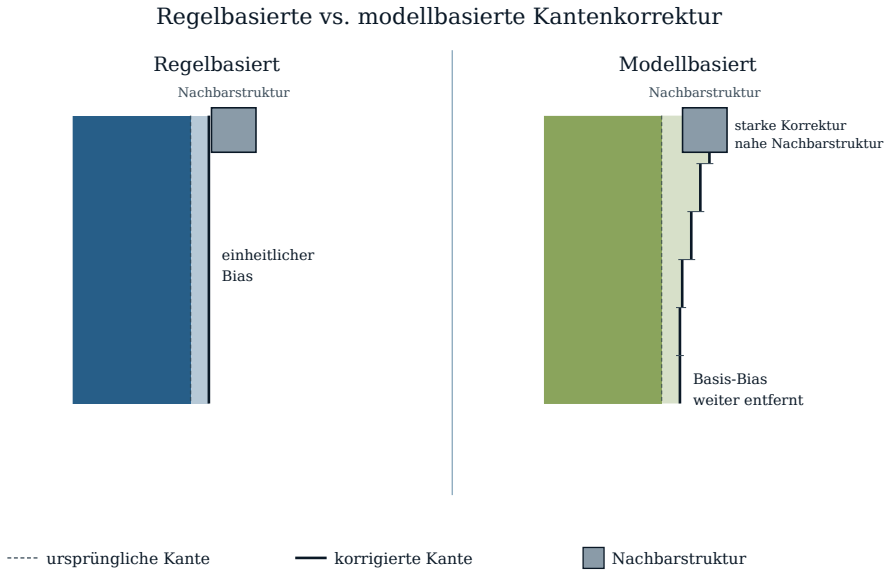
Modellbasierte OPC geht einen grundlegend anderen Weg. Statt Tabellenwerte nachzuschlagen, nutzt sie ein kalibriertes numerisches Modell des gesamten Abbildungsprozesses – optisches Modell und Resist-Modell kombiniert – um für eine gegebene Maskengeometrie den tatsächlich zu erwartenden gedruckten

Konturverlauf vorherzusagen. Jede Kante eines Polygons wird dafür in viele kurze Segmente unterteilt, die sich anschließend unabhängig voneinander verschieben lassen.

Der Korrekturalgorithmus arbeitet iterativ: Er simuliert den Druck der aktuellen Maskengeometrie, vergleicht das Ergebnis mit der Ziel-Kontur, verschiebt jedes Segment einzeln in Richtung einer besseren Übereinstimmung, und wiederholt diesen Zyklus, bis die simulierte Kontur innerhalb einer festgelegten Toleranz mit dem Soll übereinstimmt oder eine maximale Iterationszahl erreicht ist. Weil jedes Segment einzeln reagiert, kann die Korrektur in der Nähe einer komplexen 2D-Situation – etwa nahe einer Ecke – deutlich stärker ausfallen als weiter entfernt auf derselben Kante, wo einfache regelbasierte Korrektur nur einen einzigen, gemittelten Wert kennen würde.

Dieser Detailgrad hat seinen Preis: Modellbasierte OPC erfordert für jede Iteration eine Lithografiesimulation an tausenden Segmenten gleichzeitig, chipweit – ein Rechenaufwand, der um Größenordnungen über dem der regelbasierten Variante liegt. In der Praxis kombinieren moderne OPC-Flows daher beide Ansätze: Regelbasierte Korrektur für einfache, gut charakterisierte Situationen, modellbasierte Verfeinerung dort, wo die Geometrie es erfordert.

Das folgende Diagramm vergleicht beide Ansätze an derselben Kante in der Nähe einer benachbarten Struktur: links die einheitliche Verschiebung der regelbasierten Korrektur, rechts die segmentweise, an die lokale Geometrie angepasste Verschiebung der modellbasierten Korrektur.



Verifikation: Lithography Simulation & Process Window

Warum Korrektur allein nicht genügt

Eine angewendete OPC-Korrektur ist zunächst nur eine Behauptung: Das Korrekturmodell sagt voraus, dass die angepasste Maskengeometrie unter den angenommenen Prozessbedingungen das gewünschte Ergebnis druckt. Ob diese Vorhersage über das gesamte, oft milliardenfach wiederholte Chipdesign hinweg tatsächlich zutrifft – und ob sie auch unter realistischen Schwankungen im Fertigungsprozess Bestand hat –, muss anschließend gesondert geprüft werden. Diese OPC-Verifikation läuft technisch ähnlich wie die eigentliche Korrektur: eine vollflächige Lithografiesimulation, diesmal jedoch nicht zur Anpassung der Geometrie, sondern zur Kontrolle des bereits korrigierten Layouts.

Das Prozessfenster: mehr als nur der Idealfall

Belichtungsanlagen halten Fokus und Dosis nicht perfekt konstant – von Wafer zu Wafer, über die Waferfläche hinweg und selbst innerhalb eines einzelnen Belichtungsfeldes schwankt beides in einem gewissen Rahmen. Eine Korrektur, die nur beim exakten Sollfokus und der exakten Solldosis funktioniert, wäre in der Praxis wertlos. Die OPC-Verifikation simuliert deshalb nicht nur den nominalen Arbeitspunkt, sondern ein ganzes Prozessfenster aus Fokus- und Dosis-Kombinationen – typischerweise die vier Extremecken plus mehrere Zwischenpunkte – und prüft, ob die Struktur an jedem einzelnen dieser Punkte noch innerhalb der zulässigen Toleranz bleibt.

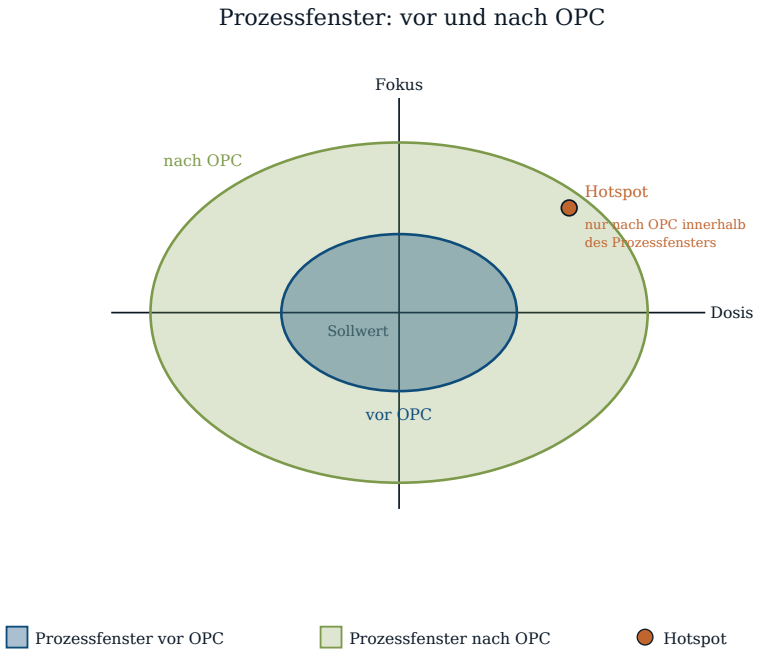
Gut ausgeführte OPC vergrößert dieses Prozessfenster spürbar gegenüber dem unkorrigierten Layout: Strukturen, die vorher nur in einem engen Fokus-Dosis-Bereich sauber druckten, bleiben nach der Korrektur auch bei größeren Abweichungen noch innerhalb der Spezifikation. Das größere Fenster ist damit ein direktes, messbares Qualitätsmaß für die Korrektur selbst – zusätzlich zur reinen Übereinstimmung am nominalen Arbeitspunkt.

Hotspots: wo die Korrektur nicht ausreicht

Stellen, an denen die simulierte Kontur an einem oder mehreren Punkten des Prozessfensters eine Design Rule verletzt – etwa weil sich zwei benachbarte Leitungen bei ungünstigem Fokus fast berühren (Bridging) oder eine Leitung sich so weit verschmälert, dass sie beinahe unterbrochen wird (Necking) –, werden als Hotspots markiert. Jeder gefundene Hotspot muss vor dem Tape-out behoben werden: durch eine gezielte lokale Nachkorrektur, notfalls durch eine manuelle Layoutänderung an dieser Stelle. Anschließend läuft die Verifikation

an der betroffenen Stelle erneut, bis auch der ungünstigste Punkt im Prozessfenster innerhalb der Toleranz bleibt.

Das folgende Diagramm zeigt das Prinzip an einem Prozessfenster-Diagramm: Auf den Achsen liegen Fokus- und Dosisabweichung vom Sollwert, die Fläche markiert die Kombinationen, bei denen die Struktur noch korrekt druckt – vor und nach OPC im Vergleich, mit einem Hotspot, der erst durch die Korrektur innerhalb des Fensters liegt.



Grenzen und Ausblick

Der Rechenaufwand von Full-Chip-OPC

Ein moderner Chip enthält mehrere Milliarden Polygone. Wird jede Kante davon für die modellbasierte Korrektur in Dutzende Segmente unterteilt und jedes Segment über mehrere Iterationen hinweg simuliert, ergibt sich ein Rechenaufwand, der selbst für große Rechenzentren eine ernsthafte Herausforderung darstellt. Full-Chip-OPC-Läufe verteilen sich heute routinemäßig über tausende CPU-Kerne parallel und benötigen trotzdem oft viele Stunden bis einzelne Tage Rechenzeit – pro Maskenebene, und ein moderner Prozess hat davon mehrere Dutzend.

Diese Datenmenge hat auch das Maskendatenformat selbst verändert: Das klassische GDSII-Format, ursprünglich für vergleichsweise einfache, unkorrigierte Layouts entworfen, stößt bei den stark fragmentierten, serifenreichen Geometrien nach OPC an praktische Grenzen. Das OASIS-Format

wurde gezielt entwickelt, um die nach der Korrektur stark vergrößerten Datenmengen kompakter zu speichern, unter anderem durch effizientere Kompression sich wiederholender Geometriemuster.

EUV: neue Wellenlänge, neue Effekte

Mit dem Übergang zur EUV-Lithografie (Extreme Ultraviolet, 13,5 nm Wellenlänge) verschiebt sich die Ausgangslage grundlegend: Da die Wellenlänge nun weit unterhalb der zu druckenden Strukturgröße liegt, ist der in Abschnitt 1 beschriebene k1-Faktor für viele Strukturen deutlich entspannter als bei der 193-nm-Immersionslithografie – die klassischen Proximity-Effekte treten in vielen Fällen schwächer auf, und der Korrekturbedarf durch Mask Bias und Serifs sinkt entsprechend.

Dafür bringt EUV eigene, neuartige Effekte mit, die eine eigene Korrekturbehandlung benötigen. Da bei dieser Wellenlänge keine transmittierenden Linsen mehr existieren, arbeitet EUV-Lithografie ausschließlich mit reflektiven Spiegeloptiken, und das Licht trifft die Maske nicht senkrecht, sondern in einem Winkel – dadurch entstehen sogenannte Mask-3D-Effekte, bei denen die endliche Dicke der Maskenabsorberschicht selbst Schatten wirft und die Abbildung asymmetrisch verzerrt, je nach Orientierung einer Struktur zum Einfallswinkel. Hinzu kommt eine geringere Photonenzahl pro Belichtung, die zu statistischen Schwankungen führt – sogenannten stochastischen Effekten –, welche mit klassischer, deterministischer OPC allein nicht vollständig beherrschbar sind.

OPC verschwindet mit EUV also nicht, sondern wandelt sich: Die in diesem Kapitel vorgestellten Grundprinzipien – Mask Bias, Serifs, SRAFs, modellbasierte Segmentkorrektur, Prozessfenster-Verifikation – bleiben gültig, werden jedoch um zusätzliche Modellkomponenten für Mask-3D- und stochastische Effekte erweitert. Bei Strukturen, die selbst mit EUV die Auflösungsgrenze einer einzelnen Belichtung unterschreiten, kommt zudem Multi-Patterning hinzu: Ein Layout wird auf mehrere Masken aufgeteilt, deren OPC-Korrekturen aufeinander abgestimmt werden müssen – ein Thema, das den Rahmen dieses Kapitels sprengt, aber auf denselben hier vorgestellten Grundlagen aufbaut.