MONOLITHISCHE 3D-SCHALTKREISE – TECHNOLOGIE UND TRENDS

Der Ausweg aus dem Skalierungs-Dilemma?

Räumliche IC-Strukturen stellen kurze Signalwege, einen geringen Energiebedarf sowie breitbandige Datenbusse bei hochintegrierten Halbleitern in Aussicht. Die Hersteller von NAND-Flash-Bausteinen haben sich als Vorreiter dieser Technologie versucht.

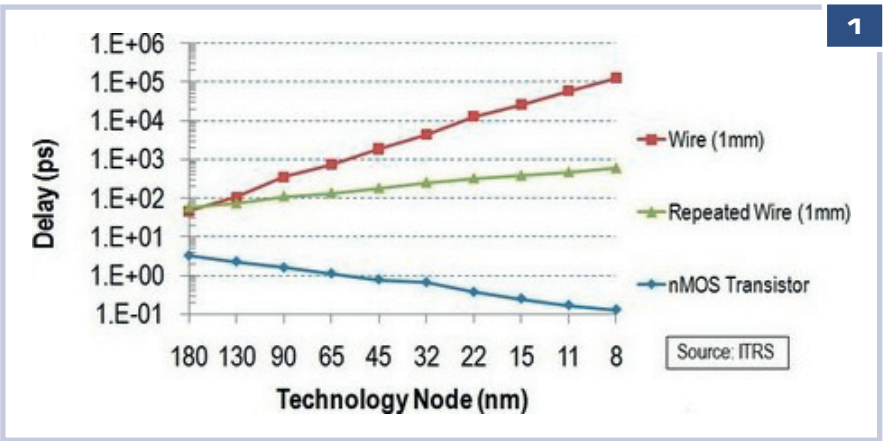


Bild 1. Verzögerungen in der Signallaufzeit (Verbindungen im Verhältnis zu Transistoren) je nach Prozessknoten

ÜBERSETZUNG UND BEARBEITUNG: HENNING WRIEDT

Mit jedem neuen Technologieknoten der Halbleiterindustrie schrumpfen die IC-Strukturen um einen Faktor von 0,7; doch das Herunterskalieren zweidimensionaler Schaltungen ist immer schwieriger geworden. Eines der schwerwiegendsten Probleme ist die Diskrepanz zwischen der schnell wachsenden Transistor-Performance und der stagnierenden Geschwindigkeit der Interconnects. Betrug die Verzögerung der Signallaufzeit einer repräsentativen 1-mm-Verbindung – im Vergleich zu der eines Transistors – noch ungefähr das Zehnfache beim 180-nm-Prozessknoten, so handelt es sich beim 32-nm-Knoten bereits um einen Faktor von 10.000.

Eine Lösung des Problems versprechen monolithische Schaltungen in drei Dimensionen, wie sie das US-amerikanische Unternehmen MonolithIC 3D propagiert. Sie stellen deutlich kürzere Signalwege, einen geringeren Energiebedarf sowie breitbandige Datenbusse bei hochintegrierten Halbleitern in Aussicht. Aber was sind die Hintergründe der 3D-Technologie und wie ist der aktuelle Stand der Entwicklungen?

Warum überhaupt 3D?

In den vergangenen 30 Jahren konnte die Halbleiterindustrie einen exponentiellen Anstieg beim Funktionsumfang und bei der Performance integrierter Schaltungen vorweisen. Die Basis dafür war erster Linie die herkömmliche Skalierung, die alle zwei Jahre eine Verbesserung der betreffenden Kennwerte um 30 Prozent mit sich brachte. Die Strukturbreiten haben sich zwischen den Jahren 1984 und 2010 von 2 µm auf 32 nm verkleinert – also fast um den Faktor 150. Künftig erwarten uns allerdings zahlreiche Schwierigkeiten, die so manchen Industrie-Insider dazu veranlassen, die Zukunft der herkömmlichen Skalierung infrage zu stellen [1, 2].

Zwei Herausforderungen des herkömmlichen Skalierens sind dabei besonders wichtig:

- Während sich die Transistor-Performance bei kleineren Geometrien verbessert, verschlechtert sich die Performance der Verbindungen. **Bild 1** verdeutlicht, dass die Signalverzögerungen eines repräsentativen 1-mm-Drahts im Vergleich zu denen eines Transistors beim 180-nm-Prozessknoten 10-mal, beim 32-nm-Prozessknoten jedoch schon 10.000-mal größer sind. Es ist möglich, die Verzögerungen der Verbindungen mithilfe von Repeatern (aus

LITERATUR

1 Scaling is dead, says IBM CTO; www.eetimes.com/electronics-news/4048782/Scaling-dead-at-130-nm-says-IBM-technologist
2 Bill Dally: Life after Moore's Law; www.forbes.com/2010/04/29/moores-law-computing-processing-opinions-contributors-bill-dally.html

22nm node	2D-IC	3D-IC 2 Device Layers	Comments
Frequency	600MHz	600MHz	
Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire capacitance to drive
Die Size (active silicon area)	50mm²	24mm²	3D-IC → **footprint 12mm²**
Power	Logic = 0.21W	Logic = 0.1W	Due to smaller Gate Size
	Reps. = 0.17W	Reps. = 0.04W	Due to shorter wires
	Wires = 0.87W	Wires = 0.44W	Due to shorter wires
	Clock = 0.33W	Clock = 0.19W	Due to less wire capacitance to drive
	Total = 1.6W	**Total = 0.8W**	

3D with 2 device layers→ 2x power reduction, ~2x active silicon area reduction vs. 2D

Tabelle A. Vergleich eines monolithischen 3D-ICs mit einem 2D-IC am gleichen Technologieknoten

Transistoren) zu reduzieren. Aber das geht auf Kosten der Siliziumfläche und erhöht den Energiebedarf. Aber selbst dann beträgt die Diskrepanz in der Signallaufzeit noch das 500-Fache am 32-nm Node.

■ Reduzierte Strukturbreiten bedeuten immens höhere Kosten bei den Halbleiterfabriken. Während eine Waferfab für die 2-µm-Technologie etwa 400 Millionen US-Dolar kostete; sind für eine 32-nm-Fabrik etwa 4 Milliarden US-Dollar fällig. Fertigungslinien mit herkömmlicher Skalierung sind inzwischen für die meisten Halbleiterunternehmen zu teuer geworden.

Was macht MonolithIC 3D?

Das kalifornische Unternehmen MonolithIC 3D hat eine Fertigungstechnik entwickelt, um monolithische 3D-ICs mit mehreren gestapelten Transistorebenen und einer besonders engen vertikalen Verschaltung herzustellen. Damit lassen sich zum Beispiel vier gestapelte Transistorebenen mit ihren vertikalen Verbindungen auf nur 50 nm realisieren.

Die Fertigung von Schaltungen mit dreidimensional gestapelten Transistoren bringt kürzere Verdrahtungen mit sich. Damit lässt sich unter anderem das oben diskutierte Problem der Signallaufzeit verbessern. Ein weiterer Vorteil ist, dass die gestapelten Ebenen mehr Transistoren pro Flächeneinheit bedeuten, ohne dass eine kostenträchtige Reduzierung der Strukturgeometrien nötig wäre.

Die Verkürzung der Leitungslängen bedeutet auch, dass sich die Logikgate-Treiber für diese Leitungen verkleinern lassen und damit werden auch die Entfernungen zwischen den Logikgates – was man als positiven Feedbackeffekt bezeichnen könnte, der letztlich die gesamte notwendige Siliziumfläche erheblich reduziert.

Analysiert wurden die Auswirkungen der Monolithic-3D Technologie mithilfe der 3D-Version es CAD-Tools „IntSim“, das sich für die Optimierung von Interconnects eignet. In der Studie diente ein 600-MHz-2D-Logikcore mit geringer Stromaufnahme, der in 22-nm-Technologie gefertigt wurde, als Maßstab. **Tabelle A** zeigt, dass ein monolithischer 3D-IC mit zwei Elementeebenen – im Vergleich zu einer 2D-IC Implementierung – eine Reduzierung beim Leistungsbedarf sowie bei der Gesamt-Siliziumfläche um die Hälfte sowie bei der benötigten Chipfläche auf ein Viertel erzielt.

Damit wird deutlich, dass monolithische 3D-ICs mit zwei Elementeebenen ähnliche Vorteile bieten können wie eine neue Technologiegeneration des herkömmlichen Skalierens (**Bild 2**). So wie die herkömmliche Skalierung mit jeder Generation die Strukturgeometrien reduziert (alle zwei Jahre um 30 Prozent), öffnen die monolithischen 3D-ICs durch ein ein-, zwei- und mehrmaliges Falten über viele Jahre hinweg den Weg zu einem kontinuierlichen Skalieren, ohne dass die Geometrien notwendigerweise reduziert werden müssen. Das ist ein besonders attraktiver Punkt, da neuzeitliche Fertigungsanlagen für das herkömmliche Skalieren, wie oben beschrieben, bereits einige Milliarden US-Dollar kosten und die Prozessentwicklung eine weitere Milliarde erfordert. Mit dem monolithischen 3D-Falten kann man die Vorteile des Skalierens dagegen weitgehend mit den vorhandenen Anlagen erreichen.

Der Weg zu MonolithIC 3D

MonolithIC 3D bietet mit Bausteinen, die in der Halbleiterindustrie sehr wohl bekannt sind, mehrere Wege zu Silizium-basierten, monolithischen 3D-Schaltungen an. Alle diese Wege sind patentiert oder zum Patent angemeldet und verwenden eine Technik, die „Ion Cut“ oder „Ebenentransfer“ genannt wird, um monokristallines Silizium auf Kupferleitungen bei weniger als 400 °C aufzubringen. Ion-Cut ist heute das vorherrschende Verfahren bei der Herstellung von Sol-Wafern (Silicon-on-Insulator).

■ **1. Weg:** Transistorherstellung oberhalb der Kupferleitungen unterhalb von 400 °C. Versenkte Kanal-Transistoren (RCATs - Recessed-ChAnnel Transistors) bilden eine Transistorfamilie, die sich mit den Verfahren von MonolithIC 3D bei weniger als 400 °C erstellen lassen. Sie werden in der DRAM-Fertigung seit dem 90-nm-Knoten verwendet. MonolithIC 3D empfiehlt diese RCATs für Logikapplikationen, sofern Technologien des Unternehmens zum Einsatz kommen sollen.

Experimentelle Daten von DRAM-Herstellern weisen darauf hin, dass RCATs im Vergleich zu Planartransistoren ähnliche Treiberströme, aber wesentlich geringere Leckage aufweisen – allerdings auf Kos-

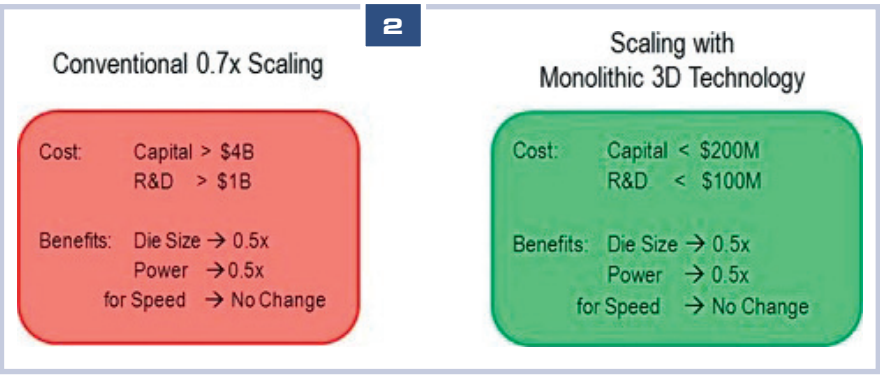
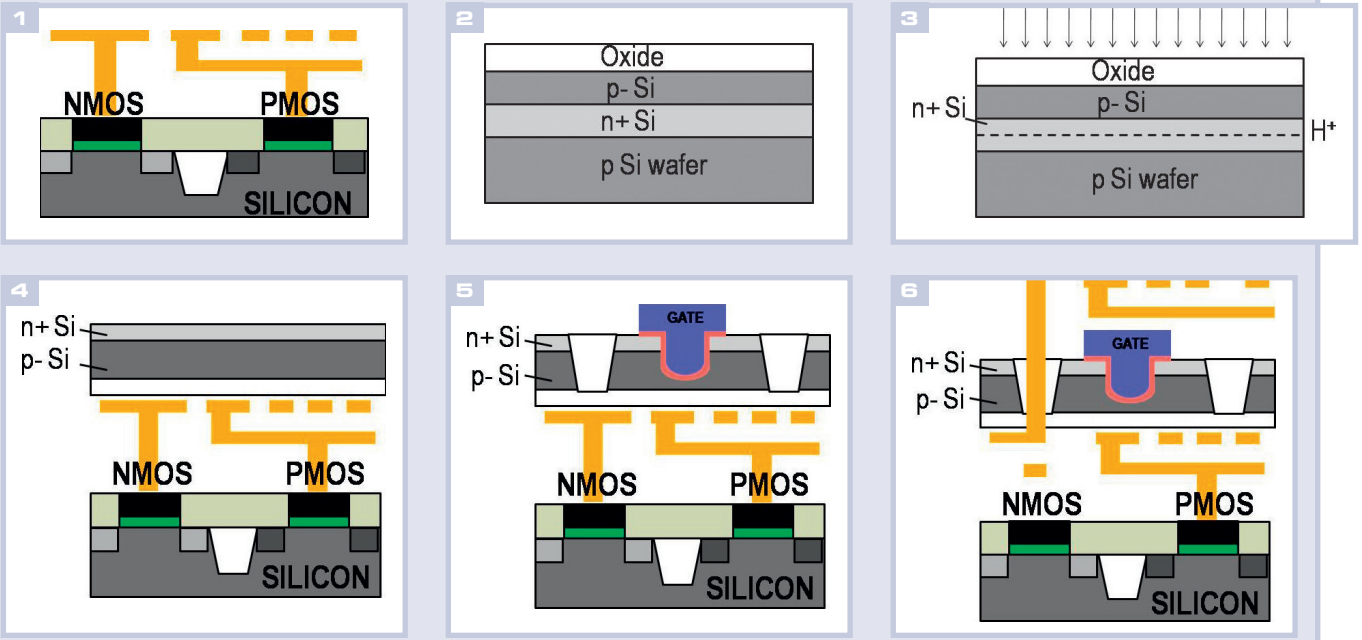
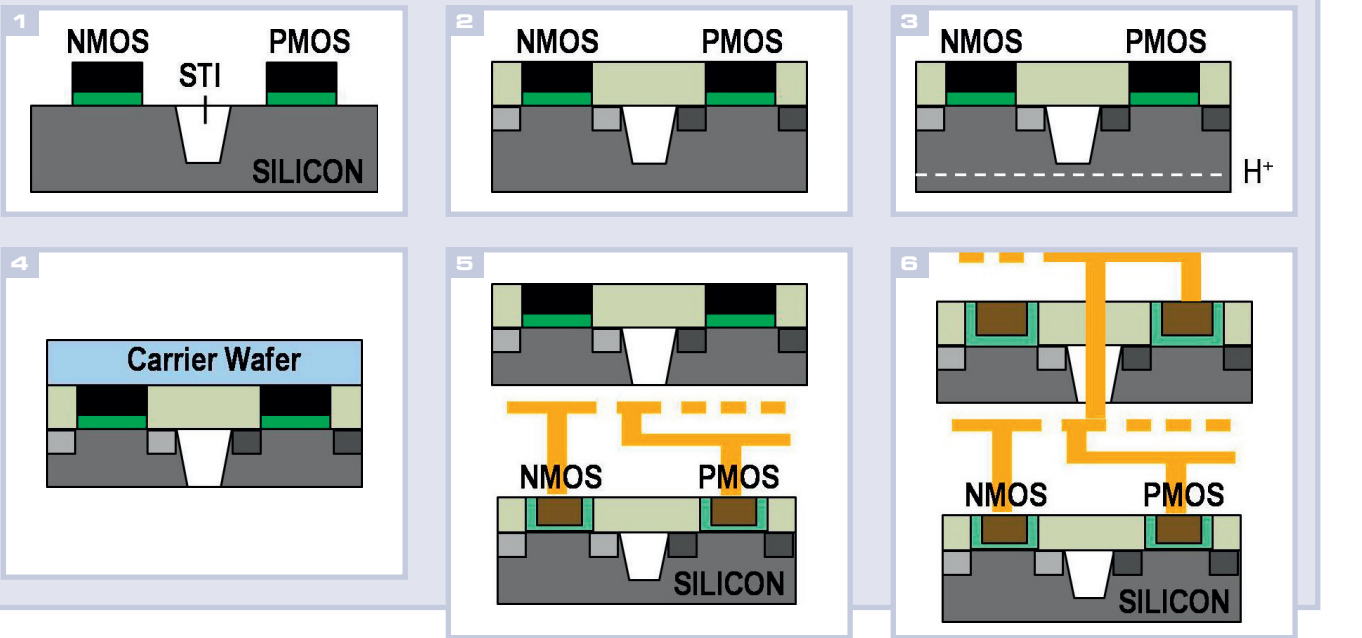


Bild 2. Herkömmliches und Monolithic-3D-Skalieren für Logikcores: Kosten und Vorteile im Vergleich

Vereinfachter Prozess für die Herstellung von RCATs oberhalb von Kupferleitungen. Schritt 1: Der Basis-Wafer des 3D-Stapels mit Transistoren und Kupferleitungen wird wie üblich hergestellt. Schritt 2: Ein Zweiebenen-Stapel mit p^--Si und n^+-Si auf einem neuen Wafer wird mit Implantier- und Epitaxial-Prozessen gefertigt. Die Dotierungen werden mit normalen Hochtemperaturtechniken aktiviert. Schritt 3: Wasserstoff wird in den Wafer mit p^-- und n^+-Siliziumregionen implantiert. Das erzeugt eine Ebene mit Defekten in der gewünschten Wafertiefe. Schritt 4: Der Wafer von Schritt 3 wird umgedreht und Oxid-zu-Oxid auf die Struktur aus Schritt 1 gebondet. Die Struktur wird dann auf der Wasserstoffebene mit mechanischen Kräften oder einer Ausglühung unter 400 °C geteilt, wodurch die dünne, dotierte Ebene des monokristallinen Siliziums auf dem Basiswafer verbleibt. Schritt 5: Die versenkten RCATs werden durch Ätzen, Deponieren und andere Verfahren geformt, die typischerweise unter 400 °C ablaufen. Das Gate-Dielektrikum und die Gate-Elektrode werden mit der ALD (Atomic Layer Deposition) aufgebracht. Schritt 6: Die elektrischen Verbindungen (Intra- und Interebenen) werden mit standardisierter BEOL-Auflösung und -Justierung erstellt, und zwar mithilfe der obersten Bauelementeebene, die etwa wenige zehn bis hundert Nanometer dünn sein kann.



Vereinfachter Prozessverlauf für das neueste Transistor-3D-Stapeln mit der Replacement-Gate-Technik. Schritt 1: Die Transistorregionen werden mit Gate-Dielektrika und Dummy-Gates erstellt. Schritt 2: Die Source-Drain-Bereiche werden mit normalen Hochtemperatur-Verfahren aufgebaut. Schritt 3: Für das Ion-Cut wird dann der Wasserstoff in der gewünschten Tiefe implantiert. Schritt 4: Der Wafer wird mit der Vorderseite nach oben mit einem vorläufigen Trägerwafer gebondet und dann in seiner Wasserstoffebene aufgetrennt. Es folgt die Polierung der exponierten Rückseite. Schritt 5: Die Wafer-Rückseite wird mit Oxid beschichtet und mit dem Basis-Wafer gebondet, der die vorgefertigten Transistoren und Verbindungen aufweist. Mit dem in Bild 3 beschriebenen besonderen Justierungen erhält man gut abgestimmte Verbindungen (Sub 50-nm) durch das Silizium. Nach dem Bonden wird der Trägerwafer entfernt. Schritt 6: Der neueste Replacement-Gate-Prozess verläuft dann normal (Entfernen und Ersetzen der Dummy-Gates). Das standardisierte BEOL folgt dann, wie gezeigt, mit der Ausbildung der Intra- und Interebenen-Verbindungen. Es ist dabei zu beachten, dass diese Prozesssequenz die Verschlechterung der Bauelemente-Eigenschaften im Hinblick auf die Implantierungsschäden am Gate-Dielektrikum (Silizium-Dioxid) in Schritt 3 verhindert.



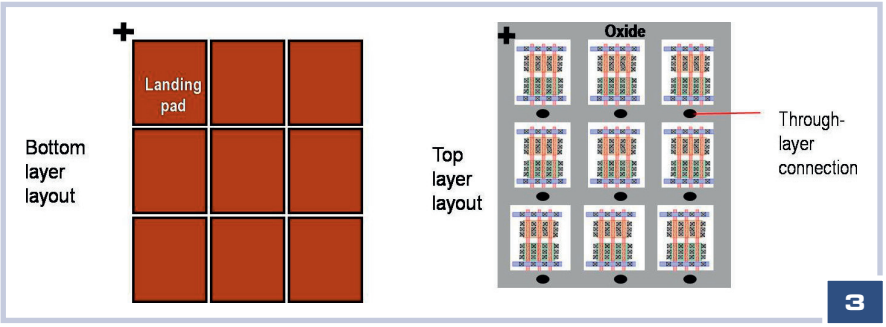


Bild 3. Sich wiederholende Layoutstrukturen für Korrekturen der Fehljustierungen beim Waferbonden: Basisebene (links) und oberste Ebene des 3D-Stapels (rechts)

ten höherer Gate-Kapazitäten. Die sechs Herstellungsschritte von RCATs in übereinander gestapelten Ebenen zeigt der obere Teil des **ⓘ-Kastens**. Dieser Prozessablauf hat drei besondere Merkmale: (1) Die Prozessschritte für die Transistoren, die weniger als 400 °C benötigen, wie die Implantat-Aktivierung, werden vor dem Ebenentransfer ausgeführt. Das stellt die Kompatibilität mit den Kupferverbindungen des Basiswafers sicher. (2) Verwendet werden versenkte Kanaltransistoren, die eine Transistorbildung mit Prozessen unter 400 °C (Ätzen, Abscheidung) zulassen. (3) Die Ebenen-zu-Ebenen Verbindungen für den 3D-Stapel werden mit fast voller Lithografie-Auflösung und -Anpassung erzielt.

■ **2. Weg:** Neueste Transistorstapelung mit der Replacement-Gate-Technologie. Ein zweiter Weg zu monolithischen 3D-ICs stellt den Stapel mit jedem neuesten Transistortyp her, der mit einem Replacement-Gate-Prozess hergestellt wird. Innovative Abgleichverfahren, in Kombination mit sich wiederholenden Layouts, erreichen im Sub-50-nm-Bereich durch das Silizium hochdichte elektrische Verbindungen. Die sechs Schritte im unteren Teil des **ⓘ-Kastens** illustrieren den Prozessablauf. Das in Schritt 5 verwendete Abgleichverfahren wird detaillierter in **Bild 3** beschrieben. Der linke Bildteil zeigt das Layout der ersten Ebene des 3D-Stapels, und rechts erkennt man das Layout der obersten Ebene. Diese oberste Ebene hat ein sich wiederholendes Layout, das einer Gate-Struktur ähnelt. Viele Unternehmen wenden sich wegen lithografischer Limitierungen ähnlichen Layouts zu. Die unterste Ebene verfügt über die so genannten Landing Pads. Deren Strukturen haben in etwa die gleiche Größe wie jede der sich wiederholenden Zellen.

Auch wenn während des Bondens ein Justierfehler auftritt, wird die angestrebte Schaltung der obersten Ebene immer noch mit einem spezifischen Landing Pad verbunden, weil die oberste Ebene aus einem Satz von sich wiederholenden Zellen besteht. Diese Toleranz gegenüber Justierfehlern ermöglicht eine hohe Dichte der Verbindungen durch die Ebenen eines Stapels.

KONTAKT

MonolithIC 3D, Inc.,
USA-San Jose, CA 95124,
Tel. 001 408 3727101,
Fax 001 408 9122303,
www.monolithic3d.com

Der Hauptvorteil dieses Verfahrens ist seine Anwendbarkeit mit jedem aktuellen Transistor, der mit einem Replacement-Gate-Prozess hergestellt wurde. Das Justierverfahren, das Monolithic 3D kommerziell implementieren will, ist ausgefeilter und flächeneffizienter als der in **Bild 3** vorgestellte Prozess. Es wird in der US Patent Application 12/847,911 beschrieben und in **Bild 4** zusammengefasst. Die braune, vertikale Markierung im Bild befindet sich auf dem Basiswafer, während die blaue, horizontale Markierung auf dem obersten Wafer zu erkennen ist. Die Vias dazwischen haben einen X-Abgleich mit der Abgleichmarkierung auf dem Basis-Wafer und einen Y-Abgleich mit der Abgleichmarkierung des obersten Wafers.

Applikationen für MonolithIC 3D

Man erwartet von tragbaren Geräten, wie Smartphones und Tablet-Computern, dass sie die Adaptierung monolithischer 3D-ICs vorantreiben. Die Power- und Platzeinsparungen des neuen Verfahrens können die Akkulaufzeit und den Formfaktor dieser Geräte verbessern. Außerdem kann die Wärmeableitung für diese Chips trotz höherer Leistungsdichten mit dem Stapeln besser gelöst werden, da die Leistungsaufnahme dieser Applikationen bis unter 1 W tendiert. Da man zudem davon ausgeht, dass die Handgeräte die Treiber für Wachstum und Volumen der Halbleiterindustrie für die nächsten zehn Jahre sein werden, dürften monolithische 3D-ICs in den kommenden Jahren an Schwungkraft gewinnen.

Monolithic 3D hat viele Applikationen für unterschiedliche Segmente der Halbleiterindustrie definiert, dazu gehören FPGA, Logik mit RCATs, Logik mit Replacement-Gate-Technik, DRAMs, NAND-Flash-Speicher, Bildsensoren und Mikrodisplays.

Eine genauere Beschreibung der Unterschiede zwischen der MonolithIC-3D- und der TSV-3D-Technologie (Through-Silicon Via) findet sich hier. *(ml)*

DER AUTOR

HENNING WRIEDT berichtet als USA-Korrespondent für EL-info und EL-info.de.

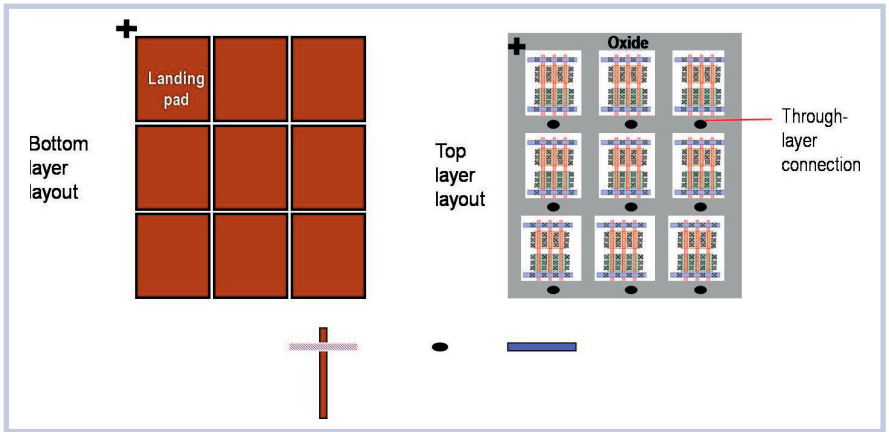


Bild 4. Ein ausgefeilteres Justierverfahren gegenüber dem in Bild 3 dargestellten Vorgang

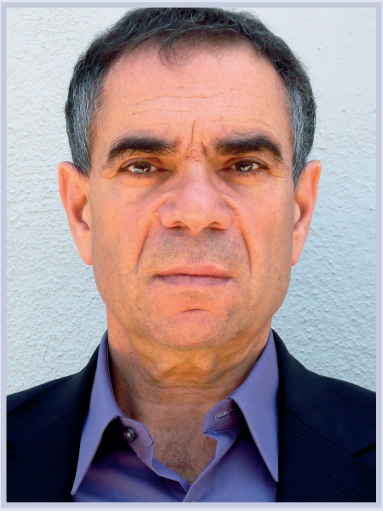
STANDPUNKT

Monolithische 3D-ICs gehen in die Serienfertigung. Zvi Or-Bach, President und CEO von MonolithIC 3D, kommentiert für EL-info den technologischen Ansatz und den Entwicklungsstand seines Unternehmens.

Herr Or-Bach, mit mehr als 100 Patenten und mit der Gründung von Start-ups wie eASIC, Chip Express und Monolithic 3D arbeiten Sie seit mehr als 35 Jahren an der vordersten Front der Halbleiterindustrie. Wie schätzen Sie diese sich immer noch entwickelnde Industrie ein?
Or-Bach: Es ist eine sehr konservative Industrie, die laufend radikale Änderungen adaptiert. Es ist eine Industrie, die erhebliche Vorteile mit der Umsetzung von Sand in Materialien bietet, die wertvoller als Gold sind. Diese Industrie verändert sich schnell, und um an ihr Teil zu haben, muss man die Gründe dieser Veränderungen verstehen und in allen Operationen integrieren.

Zurzeit scheint es so, dass die Halbleiterindustrie mit ihrer teuren Skalierungstechnologie gegen die Wand laufen wird. Sind 3D-Chips für eine gewisse Zeit die Lösung?
Or-Bach: Ja, MonolithIC 3D ist derzeit der einzige effektive Weg für die kommenden zehn oder zwanzig Jahre. Man sollte beachten, dass monolithische 3D-ICs auch sequenzielle 3D-ICs genannt werden und sich von den anderen 3D-Versionen (FinFET, TSV....) sehr stark unterscheiden.

Welches sind die großen Herausforderungen bei der Herstellung monolithischer 3D-Chips, und welches sind die wichtigsten Vorteile?
Or-Bach: Die größte Herausforderung ist das Vermeiden von Schäden oder Verschlechterungen bei der Basisschaltung mit ihren Verbindungen, die höheren Temperaturen als 400 °C nicht ausgesetzt werden sollten. Die wichtigsten Vorteile sind eine Verdopplung der Transistoren im Endprodukt bei reduzierten Kosten, die Verdopplung der Transistoren im Endprodukt ohne Erhöhung des Strombedarfs, das Überwinden des Verbindungsproblems zwischen Speicher und Logik, des so genannten Speicherwalls sowie die heterogene Integration. (Mehr dazu vermittelt der Beitrag „The Monolithic 3D Advantage“ zur IEEE 3DIC Conference vom Oktober 2013, Anm. d. Red.)



Zvi Or-Bach, President und CEO von MonolithIC 3D: Man sollte wissen, dass unsere Technologie nunmehr praktikabel ist und laufend einfacher wird – so wie sie vor wenigen Jahren als unmöglich angesehen wurde

Was müssen Chipdesigner und Kunden über 3D-Chips wissen?
Or-Bach: Sie sollten wissen, dass diese Technologie nunmehr praktikabel ist und laufend einfacher wird – so wie sie vor wenigen Jahren als unmöglich angesehen wurde. (Mehr dazu im Artikel „Scaling Makes Monolithic 3D IC Practical“, Anm. d. Red.)

Samsung hat die Serienfertigung eines 3D-V-NAND Flashspeichers angekündigt. Ist das der Ritterschlag für Ihre Technologie?
Or-Bach: Es ist für uns schlicht die Bestätigung dafür, dass monolithische 3D-ICs praktikabel und vorteilhaft sind. (Empfohlene Links dazu sowie zu aktuellen Ankündigungen von Qualcomm und CEA Leti zum Thema monolithische 3D-Logik, Anm. d. Red.)

Danke für das Gespräch.

FAZIT

Turmbau in Nanometern. Ein ernstes Problem beim Herunterskalieren von Chipstrukturen ist die Diskrepanz zwischen der schnell wachsenden Transistor-Performance und der stagnierenden Geschwindigkeit der Verbindungsleitungen. Beim 32-nm-Technologieknoten beläuft sich die Verzögerung der Signallaufzeit einer repräsentativen 1-mm-Verbindung – im Vergleich zu der eines Transistors – bereits auf das 10.000-Fache. Eine Lösung des Problems versprechen monolithische 3D-Schaltungen. Sie stellen deutlich kürzere Signalwege, einen geringeren Energiebedarf sowie breitbandige Datenbusse in Aussicht. Der Beitrag erklärt die Hintergründe der 3D-Technik, fasst den aktuellen Entwicklungsstand zusammen und verlinkt auf verschiedene Anwendungsbeispiele.